UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FÁBIO MANOEL FRANÇA LOBATO

ESTRATÉGIAS EVOLUCIONÁRIAS PARA OTIMIZAÇÃO NO TRATAMENTO DE DADOS AUSENTES POR IMPUTAÇÃO MÚLTIPLA DE DADOS

TD 03/2016

UFPA / ITEC / PPGEE

Campus Universitário do Guamá

Belém-Pará-Brasil

2016

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ESTRATÉGIAS EVOLUCIONÁRIAS PARA OTIMIZAÇÃO NO TRATAMENTO DE DADOS AUSENTES POR IMPUTAÇÃO MÚLTIPLA DE DADOS

FÁBIO MANOEL FRANÇA LOBATO

Tese submetida à avaliação da Banca Examinadora aprovada pelo colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará e julgada adequada para a obtenção do Grau de Doutor em Engenharia Elétrica na área de computação aplicada.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2016

Lobato, Fábio Manoel França
Estratégias evolucionárias para otimização no tratamento de dados ausentes por imputação múltipla de dados / Fábio Manoel França Lobato; orientador, Ádamo Lima de Santana. - 2016.
157 p. : il. (algumas color.) ; 30 cm.

Tese (Doutorado) – Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica/ITEC/PPGEE.

1. imputação múltipla de dados. 2. dados ausentes. 3. computação evolutiva. 4. algoritmos genéticos. 5. algoritmo genético multiobjetivo. I. Santana, Ádamo Lima de. II. Universidade Federal do Pará. III. Programa de Pós-graduação em Engenharia Elétrica. IV. Título

CDD 519.53

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ESTRATÉGIAS EVOLUCIONÁRIAS PARA OTIMIZAÇÃO NO TRATAMENTO DE DADOS AUSENTES POR IMPUTAÇÃO MÚLTIPLA DE DADOS

AUTOR: FÁBIO MANOEL FRANÇA LOBATO

TESE SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: __/__/____

BANCA EXAMINADORA:

Prof. Dr. Ádamo Lima de Santana

(ORIENTADOR - UFPA)

Prof. Dr. Roberto Célio Limão de Oliveira

(MEMBRO - ITEC - UFPA)

Profa. Dr. Adriana Rosa Garcez Castro

(MEMBRO - ITEC - UFPA)

Prof. Dr. Claudomiro de Souza de Sales Júnior

(MEMBRO - ICEN - UFPA)

Prof. Dr. André Ponce de Leon Carvalho

(MEMBRO - ICMC - USP)

Profa. Dr. Solange Oliveira Rezende

(MEMBRO - ICMC - USP)

VISTO:

Prof. Dr. Evaldo Gonçalves Pelaes

(COORDENADOR DO PPGEE/ITEC/UFPA)

AGRADECIMENTOS

Não há palavras para descrever minha gratidão à minha família por todo apoio, em especial aos meus pais, Ivan e Eluiza, agradeço pelos esforços incomensuráveis à minha educação.

Agradeço ao professor Ádamo Lima de Santana pela orientação ao longo destes sete anos. Obrigado por, desde a graduação, guiar meus passos acadêmicos e proporcionar oportunidades de aprendizado ímpares.

Agradeço também a todos os professores que, desde o colégio, vêm acreditando no meu potencial e incentivando a seguir atrás dos meus sonhos. Principalmente aos professores, ex-professores e funcionários da Faculdade de Engenharia da Computação e do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará.

Meu muito obrigado ao professor Fernando B. E. Otero, por me receber na Universidade de Kent durante o período-sanduíche, e pelo tempo dedicado à minha orientação.

Agradeço aos membros dos laboratórios que tive a honra de fazer parte, LPRAD e LINC, pela convivência fundamental para meu amadurecimento acadêmico e pessoal; aos membros do grupo de estudo de dados ausentes do LPRAD e LINC, Lilian Dias, Vincent Tadaiesky, Igor Araújo, Damares Resende, Kevin Freire e Antônio Jacob, por todo o auxílio e discussões; e também aos demais co-autores que participaram ativamente da pesquisa, Leonardo Ramos e Prof. Claudomiro Sales.

Meus agradecimentos aos colegas de trabalho da Universidade Federal do Oeste do Pará; e aos antigos colegas de trabalho da Universidade da Amazônia, principalmente ao professor Antônio Jacob pelas oportunidades concedidas e amizade dedicada.

Aos amigos, a "família" que me foi permitido escolher, por me apoiarem e proporcionarem momentos únicos e fundamentais para seguir em frente, sou muito feliz por tê-los em minha vida.

Meus sinceros agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro para o desenvolvimento deste trabalho.

"Estamos sós e sem desculpas. É o que traduzirei dizendo que o homem está condenado a ser livre. Condenado, porque não se criou a si próprio; e no entanto livre, porque uma vez lançado ao mundo é responsável por tudo quanto fizer." (Jean Paul Sartre)

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

AE Algoritmos Evolucionários

AG Algoritmos Genéticos

ARCH *autoregressive conditional heteroskedasticity*

ARIMA *autoregressive integrated moving average*

BR *Binary Relevance*

CE Computação Evolucionária

CP Conjunto de Pareto

DAC Acurácia Distributiva

ECM *Evolving Clustering Method*

EM *Exact match*

FP Fronteira de Pareto

GPMI *Genetic Programming Multiple Imputation*

HL *Hamming Loss*

IM Imputação Múltipla

ImpD Imputação de Dados

KDD *Knowledge-Discovery in Databases*

kNN *k-Nearest Neighbor*

KNNI *k-Nearest Neighbor Imputation*

LWL *Locally Weighted Learning*

MAR *Missing at random*

MCAR *Missing completely at random*

NMAR *Not Missing at Random*

NRMSE *Normalized Root Mean Square Error*

NSGA *Nondominated Sorting Genetic Algorithm*

NSGA-II Fast Nondominated Sorting Genetic Algorithm

nu-SVM *nu-Support Vector Machine*

PAC Acurácia Preditiva

POM Problemas de Otimização Multiobjetivo

RMSE *Root Mean Square Error*

rSVD *regulated Singular Value Decomposition*

SPEA *Strength Pareto Evolutionary Algorithm*

SPEA2 *improved* SPEA

TIC Tecnologia da Informação e Comunicação

TVA Tratamento de Valores Ausentes

VA Valores Ausentes

VEGA *Vector Evaluated Genetic Algoritm*

WNR *Wilson's Noise Ratio*

RESUMO

A análise de dados envolve aquisição e organização de informação com o objetivo de se obter conhecimento a partir deles, propiciando avanços científicos nos mais variados campos, bem como provendo vantagens competitivas às corporações. Neste âmbito, um problema ubíquo na área merece destaque, os valores ausentes, pois a maior parte das técnicas de análise de dados não consegue lidar de forma satisfatória com dados incompletos, impactando negativamente o resultado final. Visando contornar os efeitos danosos desta problemática, diversos trabalhos vêm sendo desenvolvidos nas áreas de análise estatística e aprendizado de máquina, com destaque para o estudo de métodos de Imputação Múltipla de Dados (IMD), que consiste no preenchimento dos dados ausentes por valores plausíveis. Tal metodologia pode ser vista como um problema de otimização combinatória, onde buscam-se valores candidatos à imputação de forma a reduzir o viés imposto por esta problemática. Meta-heurísticas, em especial, métodos baseados em Computação Evolucionária (CE) têm sido aplicadas com sucesso em problemas de otimização combinatórios. Apesar dos recentes avanços na área, percebe-se algumas falhas na modelagem dos métodos de imputação baseados em CE existentes. Visando preencher tais lacunas encontradas na literatura, esta tese apresenta uma descrição da IMD como um problema de otimização combinatória e propõe métodos baseados em CE neste contexto. Além disso, em virtude das falhas encontradas na modelagem dos métodos recentemente propostos na literatura e da necessidade de se adotar diferentes medidas de desempenho para avaliar a eficiência dos métodos de imputação, também é proposto neste projeto de tese um algoritmo genético multiobjetivo para a imputação de dados no contexto de classificação de padrões. Este método mostra-se flexível quanto aos tipos de dados, além de evitar a análise de caso completo. Dado a flexibilidade da abordagem proposta, é possível ainda utilizá-lo em outros cenários como no aprendizado não supervisionado, classificação multirrótulo e em análise de séries temporais.

PALAVRAS-CHAVES: imputação múltipla de dados. dados ausentes. computação evolutiva. algoritmos genéticos. algoritmo genético multiobjetivo.

ABSTRACT

The data analysis process includes information acquisition and organization in order to obtain knowledge from them, bringing scientific advances in various fields, as well as providing competitive advantages to corporations. In this context, an ubiquitous problem in the area deserves attention, the missing data, since most of the data analysis techniques can not deal satisfactorily with this problem, which negatively impacts the final results. In order to avoid the harmful effects of missing data, several studies have been proposed in the areas of statistical analysis and machine learning, especially the study of Multiple Data Imputation, which consists in the missing data substitution by plausible values. This methodology can be seen as a combinatorial optimization problem, where the goal is to find candidate values to substitute the missing ones in order to reduce the bias imposed by this issue. Metaheuristics, in particular, methods based in evolutionary computing have been successfully applied in combinatorial optimization problems. Despite the recent advances in this area, it is perceived some shortcomings in the modeling of imputation methods based on evolutionary computing. Aiming to fill these gaps in the literature, this thesis presents a description of multiple data imputation as a combinatorial optimization problem and proposes imputation methods based on evolutionary computing. In addition, due to the limitations found in the methods presented in the recent literature, and the necessity of adoption of different evaluation measures to assess the imputation methods performance, a multi-objective genetic algorithm for data imputation in pattern classification context is also proposed. This method proves to be flexible regarding to data types and avoid the complete-case analysis. Because the flexibility of the proposed approach, it is also possible to use it in other scenarios such as the unsupervised learning, multi-label classification and time series analysis.

KEYWORDS: multiple imputation. missing data. evolutionary computing. genetic algorithms. multi-objective genetic algorithm.

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E DESAFIOS

A análise de dados envolve a aquisição e organização da informação de forma a se obter conhecimento a partir deles, propiciando avanços científicos nos mais variados campos, bem como provendo vantagens competitivas às corporações (WITTEN; FRANK; HALL, 2011; SCHUTT; O'NEIL, 2013; HAIR, 2014). Dado sua larga aplicabilidade, há um grande interesse por esta área de estudo, sobretudo no desenvolvimento de métodos que aumentem a confiabilidade do resultado final (HAN; KAMBER, 2006; FÁVERO et al., 2009). Neste âmbito, o problema da ausência de dados merece destaque devido sua ubiquidade, aliado ao fato da maior parte dos métodos de análise não terem sido desenvolvidos para lidar satisfatoriamente com este problema (LITTLE; RUBIN, 2002; GRAHAM, 2009). Consequentemente, um dos primeiros passos do processo de análise de dados é verificar e documentar a extensão dos Valores Ausentes (VA) (SAINANI, 2015).

Diversas metodologias vêm sendo desenvolvidas com o intuito de mitigar os efeitos nocivos da ausência de dados para com a qualidade das informações extraídas (LUENGO; GARCÍA; HERRERA, 2012); sendo que abordagens baseadas em aprendizado de máquina e métodos importados da teoria de aprendizado estatístico são as mais intensamente estudadas e utilizadas nesta área (GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009). A maior parte destas metodologias são aplicadas durante o pré-processamento de dados e visam substituir os dados faltosos por valores plausíveis, estratégia conhecida por Imputação de Dados (ImpD) (LITTLE; RUBIN, 2002; GARCÍA; LUENGO; HERRERA, 2015). Esta estratégia pode ser vista como um problema de otimização, onde busca-se uma combinação de valores candidatos à imputação de forma a reduzir o viés imposto pelos VA (OLIVEIRA; COELHO, 2009).

Meta-heurísticas têm sido amplamente utilizadas como métodos de busca e otimização para resolver problemas complexos (LUKE, 2013; ROTHLAUF, 2011), com destaque para a Computação Evolucionária (CE), a qual baseia-se na Teoria da Evolução de Darwin e tem sido bem sucedida na resolução de tarefas de engenharia que vão da perspectiva do molecular ao astronômico (EIBEN; SMITH, 2003; ALBA; LUQUE; NESMACHNOW, 2013; EIBEN; SMITH, 2015). Exemplos clássicos de algoritmos pertencentes à CE são os Algoritmos Genéticos (AG), estratégias evolutivas e programação genética (BÄCK, 1996; De Castro, 2007); como mais recentes, é possível citar a evolução diferencial e a otimização por enxame de partículas (EIBEN; SMITH, 2015).

Apesar dos recentes avanços na área e dos benefícios dos métodos de imputação baseados em CE, existem desafios de pesquisa em aberto que limitam a plena utilização das aborda-

gens existentes. A seguir, são descritos os desafios de pesquisa que estão mais relacionados ao contexto Imputação Múltipla (IM) de dados baseados na estratégia evolutiva.

Formalização do Problema: posto que a imputação múltipla de dados pode ser tratada como um problema de otimização, não há na literatura uma formalização. Um desafio de pesquisa é propor uma definição formal da imputação de dados como um problema de otimização, de forma a fornecer um arcabouço que conceda maior flexibilidade e robustez às soluções nela baseadas. Com uma formalização apropriada para a ImpD, é possível reduzir o espaço de busca em bases de dados complexas (com alta dimensionalidade e grande quantidade de VA) por meio de adoção de estratégias de estratificação, uma vez que os métodos evolucionários baseiam-se no princípio de geração-e-teste (EIBEN; SMITH, 2015), o que impacta diretamente no custo computacional caso a complexidade da função de avaliação seja alta; investigar estratégias para evitar a análise de caso completo, prevenindo a perda de informação potencialmente útil contida nos exemplos incompletos (EEKHOUT et al., 2012); incorporar conhecimento de fundo por meio de restrições *Must-Link* e *Cannot-Link* (WAGSTAFF et al., 2001), evitando a imputação de valores espúrios, como associar em um mesmo exemplo valor "Homem" ao valor "Gravidez positiva" (BARALDI; ENDERS, 2010); como também facilitar a portabilidade de soluções para diferentes nichos de aplicação. Portanto, algumas falhas conceituais encontrados na literatura de ImpD baseada em computação evolucionária podem ser suplantadas caso haja uma formalização adequada do problema, sendo assim, a definição formal do processo de imputação múltipla de dados como um problema de otimização representa um importante desafio de pesquisa;

Desenvolvimento de métodos de imputação flexíveis quanto ao tipo de dados: aplicações do mundo real são geralmente compostas de conjuntos de dados com atributos categóricos e numéricos (SCHAFER, 1997), contudo, uma grande quantidade de métodos de imputação trabalham exclusivamente à um tipo de atributos apenas ou possuem restrições na exploração de atributos de diferentes tipos, como o caso dos métodos de imputação baseados em *k-Nearest Neighbor* (kNN), onde a escolha de uma medida de proximidade geralmente beneficia atributos categóricos ou numéricos (ZHANG; JIN; ZHU, 2011; Van Hulse; KHOSHGOFTAAR, 2011; ZHANG, 2012). Esta restrição não é exclusiva das abordagens baseadas em *k*NN, ela também se aplica à métodos estatísticos, os quais beneficiam atributos numéricos em detrimento dos categóricos e ordinais (LITTLE; RUBIN, 2002; GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009). Por este motivo, diversos estudos vêm sendo desenvolvidos com o intuito de desenvolver métodos de imputação baseados em CE que sejam robustos à bases que possuam atributos numéricos, categóricos ou mistos. Devido às restrições expostas, o estudo e proposição de métodos de imputação flexíveis a esses tipos de dados é um desafio de pesquisa pertinente à área desta tese de doutorado;

Métodos de imputação baseadas em CE e análise de caso completo: um item comum aos métodos de imputação baseados em CE é que eles recaem em análise de casos completos (Figueroa García; KALENATIC; López Bello, 2008; de Andrade Silva; HRUSCHKA, 2009; Figueroa García; KALENATIC; López Bello, 2010; Figueroa García; KALENATIC; López Bello, 2011; AYDILEK; ARSLAN, 2013). Ou seja, as instâncias com valores ausentes não são utilizadas para predizer os valores a serem imputados (EEKHOUT et al., 2012). Além da perda de informação potencialmente útil, a aplicação de tais métodos em cenários reais torna-se inviável, pois frequentemente encontram-se conjuntos de dados com mais de 80% de instâncias com pelo menos um atributo com VA. Sendo assim, o estudo e proposição de métodos de imputação flexíveis aos tipos de dados representa um desafio de pesquisa notório na análise de dados;

Estudo de estratégias para lidar com medidas de desempenho conflitantes: há diversas formas de avaliar o desempenho de métodos de imputação, uma das mais usuais é realizada por meio de testes utilizando conjuntos de dados artificiais, possibilitando a adoção do erro quadrático médio calculado entre o valor real e o predito pelo método de imputação. No entanto, algumas discussões acerca dessa abordagem devem ser destacadas: *i)* a indução da ausência de dados nem sempre reflete o modelo real de aleatoriedade da ausência de dados, consequentemente, um método com um bom desempenho neste tipo de cenário poderá apresentar resultados enviesados em casos reais (LITTLE; RUBIN, 2002; GRAHAM, 2009); *ii)* alguns autores sugerem que a consideração de medidas baseadas na tarefa de modelagem, como por exemplo a classificação de padrões, é imprescindível, uma vez que a tarefa de modelagem é o objetivo final do processo de análise dos dados (GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009), e ainda, alguns estudos apontam que uma melhor acurácia preditiva do método de imputação não conduz a um menor viés no processo de modelagem (HRUSCHKA et al., 2009; SILVA; HRUSCHKA, 2013); *iii)* há também discussões sobre a utilização do erro quadrático médio como única medida de desempenho, pois tal erro, sozinho, não reflete a variabilidade da amostragem e incertezas a respeito de um modelo de ausência de dados (LITTLE; RUBIN, 2002); *iv)* por fim, discute-se a proporcionalidade entre as medidas de desempenho, visto que algumas delas apresentam comportamentos conflitantes - enquanto uma otimiza, a outra decai. Deste modo, um dos desafios da literatura é estudar estratégias flexíveis à adição de diferentes medidas de desempenho, e ainda, propor métodos que lidem de forma satisfatória com medidas de desempenho conflitantes;

Análise de convergência, sintonização e controle de parâmetros: algoritmos evolutivos produzem soluções aproximadas (ROTHLAUF, 2011). Isto, aliado ao fato de serem algoritmos estocásticos, a análise de convergência a fim de avaliar a evolução e satisfatoriedade das soluções é imprescindível (DERRAC et al., 2014). Outro ponto em aberto é a sintonização de parâmetros dos algoritmos evolutivos no cenário de Tratamento de Valo-

res Ausentes (TVA), a qual pode ser considerada de dois pontos de vista: da escolha de parâmetros que otimizem a performance do método, e do estudo da dependência do desempenho em relação à parametrização (EIBEN; SMIT, 2011). Em particular, a segunda perspectiva é a mais interessante no contexto em questão, pois assim é possível extrair informações relevantes ao problema. Por meio das análises de convergência e sintonização de parâmetros também é factível se aprofundar na análise do comportamento dos algoritmos evolutivos aplicados à imputação de dados, possibilitando o uso de valores apropriados para os parâmetros nos diferentes estágios do processo de busca e até mesmo diminuir o número de parâmetros informados pelo usuário (KARAFOTIAS; HOOGENDOORN; EIBEN, 2015). Por ser uma lacuna na literatura, este tópico é um desafio de pesquisa em aberto.

Extrapolação para outras tarefas de análise de dados: o tratamento de valores ausentes tem sido discutido extensivamente na literatura de análise estatística (LITTLE; RUBIN, 1987; SCHAFER, 1997; ALLISON, 2001; LITTLE; RUBIN, 2002), atualmente observa-se uma maior tendência ao estudo dos métodos de TVA baseados em aprendizado de máquina. Sendo a classificação de padrões uma das tarefas de mineração de dados mais recorrente, percebe-se uma convergência de trabalhos envolvendo imputação de dados a este tipo de análise (GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009). Contudo, outros nichos de aplicação são ainda mais suscetíveis a incidência de valores ausentes, como também mais sensíveis, uma vez que não há técnicas de análise robustas à esta problemática. Portanto, apenas duas estratégias para mitigar os efeitos danosos dos VA podem ser empregadas neste contexto: *i)* a remoção de exemplos ou atributos com valores ausentes que consequentemente têm-se a perda de informação, o que é crítico em conjuntos de dados com grande incidência de VA; *ii)* a imputação de dados, permitindo o uso das técnicas de análise usuais. Apesar dos avanços na área (Figueroa García; KALENATIC; López Bello, 2010; HONAKER; KING; KING, 2013; JUNGER; Ponce de Leon, 2015), o estudo e desenvolvimento de métodos de imputação para outros nichos de aplicação é um desafio de pesquisa interessante e que também pode se beneficiar da definição formal da imputação de dados como problema de otimização.

1.2 OBJETIVOS

A recorrência dos valores ausentes no processo de análise de dados, os benefícios da imputação múltipla nos mais diversos domínios de aplicação e as lacunas encontradas, conforme discutido nos desafios de pesquisa, são as motivações para o desenvolvimento deste projeto de doutorado que tem como principal propósito investigar e desenvolver métodos de imputação múltipla baseados em computação evolucionária que sejam flexíveis aos tipos de dados e ao domínio de aplicação, eficientes frente a medidas de desempenho conflitantes e que reduzam o viés imposto pela ausência dos dados.

A fim de alcançar o propósito do projeto, definiram-se três objetivos que tratam de pontos de pesquisa em aberto e que pertencem ao escopo da tese:

1. Propor e testar uma definição formal para a imputação múltipla de dados como um problema de otimização, permitindo suplantar as falhas presentes nos métodos de imputação de dados baseados em CE recentemente propostos. Baseadas nesse objetivo, algumas perguntas guiam o desenvolvimento do trabalho:

    - Qual a importância de um modelo formal para imputação múltipla de dados como um problema de otimização?
    - Como definir e representar o espaço de busca e restrições de forma a respeitar as características intrínsecas da base?
    - Quais as estratégias de busca que melhor se aplicam ao modelo formal proposto?

2. Desenvolver e aprimorar algoritmos de imputação múltipla de dados baseados em CE eficientes, que considerem conjuntos de dados com atributos de tipos mistos, evitem a análise de caso completo e que lidem de forma satisfatória com medidas de desempenho conflitantes. Assim, as seguintes perguntas norteiam este trabalho:

    - Quais as vantagens e desvantagens da utilização de computação evolucionária para realizar a imputação de dados?
    - Como tratar atributos numéricos e categóricos igualitariamente e ainda, levar em consideração exemplos com valores ausentes?
    - Como avaliar soluções candidatas e combinar as soluções levando-se em consideração medidas de desempenho conflitantes?

3. Analisar o comportamento de métodos de imputação múltipla baseados em computação evolucionária em relação à convergência e parametrização, de forma a estudar estratégias de sintonização e controle de parâmetros adequadas. As seguintes perguntas orientam esta frente de trabalho:

- Qual o impacto da variação dos parâmetros no desempenho do método?
- As informações acerca da convergência e parametrização são importantes para o domínio de aplicação ou podem ajudar o especialista do domínio a melhor entender a ausência de dados na base em análise?

À luz destes objetivos principais, podem-se destacar os seguintes objetivos específicos que visam a adoção de métodos de imputação múltipla baseados em computação evolucionária em cenários reais:

1. Utilizar e validar a descrição proposta no desenvolvimento de novos algoritmos para a imputação múltipla baseados em CE, além de refiná-la em uma definição formal a fim de reduzir o espaço de busca e permitir a inclusão de conhecimento de fundo;
2. Desenvolver, testar e validar codificações de soluções apropriadas à imputação múltipla de dados, por meio de experimentos controlados – tanto em bases de *benchmarking* quanto para estudos de caso disponíveis;
3. Analisar a adoção de medidas de desempenho conflitantes a fim de se estudar estratégias para a avaliação e escolha das soluções candidatas;
4. Estudar estratégias para incorporar a informação contida nos exemplos com valores ausentes na estimação dos valores a serem imputados;
5. Adotar estratégias de análise e controle de parâmetros no cenário em estudo;
6. Extrapolar os métodos desenvolvidos para outros domínios de aplicação além da classificação de padrões, como análise de séries temporais e classificação multirrótulo;
7. Realizar uma análise crítica acerca dos modelos e métodos propostos a fim de identificar limitações a serem melhoradas.

1.3 PRINCIPAIS RESULTADOS

Baseadas nos objetivos acima descritos, algumas contribuições científicas foram realizadas:

Revisão sistemática sobre métodos de tratamento de valores ausentes: uma revisão sistemática sobre métodos para tratamento de valores ausentes foi planejada, conduzida e está reportada inicialmente neste projeto de tese. Em sua etapa de condução, 9.000 publicações foram identificadas, com 132 artigos passando pelos critérios de seleção e 40 trabalhos devidamente analisados. Como resultado, percebeu-se uma tendência clara no uso de imputação de dados como o principal método para lidar com VA, adicionalmente,

percebeu-se uma falta de padronização nos experimentos, o que dificulta a replicação, avaliação e comparação fidedigna entre os métodos recentemente propostos, seja pela academia ou indústria.

***Framework* experimental para testes envolvendo imputação de dados:** foi proposto um *framework* para a implementação de testes envolvendo métodos de imputação de dados no contexto do aprendizado supervisionado (JESUS et al., 2013). A motivação para o desenvolvimento deste *framework* foi identificada na revisão sistemática conduzida: a falta de padronização nos testes envolvendo métodos para tratamento de valores ausentes, o que dificulta a replicação e consequentemente a comparação fidedigna entre eles. Portanto, o objetivo deste trabalho é fornecer aos pesquisadores uma sequência de etapas que permitam a fácil replicação dos experimentos no contexto de classificação de padrões.

Descrição da imputação múltipla de dados como um problema de otimização: é proposto neste projeto de tese uma descrição formal para a imputação múltipla como um problema de otimização combinatória. Esta descrição cobre a identificação e especificação da imputação de dados como um problema de otimização, indicando possíveis representações das soluções candidatas, inicialização e operadores de busca. Medidas de avaliação das soluções são indicadas de acordo com o que se está estabelecido na literatura de análise estatística e aprendizado de máquina. Também são discutidos eventuais gargalos na adoção de determinadas medidas de avaliação e estratégias para reduzir o espaço de busca.

Algoritmo genético para a imputação de dados: foi proposto e desenvolvido um algoritmo genético mono-objetivo para a imputação de dados para otimizar classificadores baseados em aprendizado de máquina (LOBATO et al., 2015b). A maior parte dos métodos de imputação são restritos a um tipo de variável apenas (categóricas ou numéricas) e recaem em análise de caso completo. Portanto, o método proposto visa preencher tais lacunas na literatura, lidando de forma satisfatória com os tipos de dados supracitados, além de levar em consideração instâncias com valores ausentes. Como função de avaliação, adotou-se a acurácia do classificador, de forma a incorporar a informação da construção do modelo na escolha dos valores a serem imputados. Em um trabalho posterior, os testes foram estendidos a mais conjuntos de dados, tanto com VA existentes quanto com induzidos, avaliou-se a convergência das soluções e a sensibilidade da parametrização do algoritmo genético para a imputação de dados. Os resultados mostram que o método proposto obtém performance superior aos métodos de imputação comparados; e o comportamento do algoritmo genético desenvolvido é estudado em relação à adoção de diferentes valores para os parâmetros quantitativos.

Algoritmo genético multiobjetivo para para imputação de dados: foi proposto e desenvolvido um algoritmo genético multiobjetivo para a imputação de dados (LOBATO et al., 2015a). O algoritmo é baseado no algoritmo NSGA-II e incorpora as características do

algoritmo mono-objetivo descrito acima pois leva em consideração instâncias com VA e informação da construção do modelo de classificação, e ainda, lida com atributos mistos da mesma forma. O diferencial está na incorporação de outra medida de desempenho na função de aptidão, o erro quadrático médio obtido a partir do valor imputado e o valor real. Por consequência da adoção desta medida de avaliação, apenas conjuntos de dados com VA induzidos foram usados nos experimentos. Os resultados obtidos mostram que o método multiobjetivo proposto apresenta um bom *trade-off* para medidas de avaliação conflitantes, ademais, o método mostra-se flexível quanto ao domínio de aplicação, uma vez que a função de avaliação pode ser facilmente modificada.

Extrapolações de métodos evolucionários para diferentes domínios: neste projeto de doutorado também é proposto um método de imputação para dados advindos de séries temporais utilizando programação genética, o GPimpute. Também é proposto um método de imputação que evolui soluções obtidas por métodos de imputação simples, a qual é aplicada no cenário de classificação multirrótulo. Este último está em fase de teste a fim de se avaliar e validar os resultados obtidos. Em resultados preliminares este método mostrou-se competitivo e bastante eficaz em relação ao custo computacional.

1.4 ORGANIZAÇÃO DO TEXTO

O restante deste trabalho esta organizado como segue:

Cap. 2 - Fundamentação teórica: neste capítulo conceitos pertinentes às áreas correlacionadas com a pesquisa são apresentados. Mais especificamente, disserta-se acerca de valores ausentes, suas causas e impactos sobre os mecanismos de ausência de dados e sobre os principais paradigmas de tratamento de valores ausentes. Também apresentam-se conceitos em relação à computação evolucionária, com destaque para algoritmos genéticos, programação genética, estratégias multiobjetivo, sintonização e controle de parâmetros, e finalmente abordam-se métodos estatísticos para avaliação de algoritmos evolucionários. Por fim, são apresentadas as considerações finais.

Cap 3. - Trabalhos Correlatos: neste capítulo os trabalhos correlacionados à esta tese são discutidos. Devido a grande quantidade de estudos no tema, dividiu-se o capítulo em três seções principais. Primeiro as revisões da literatura do tema são apresentadas para então discutir os trabalhos que conduziram estudos comparativos; então discutem-se trabalhos que utilizam-se da computação evolucionária no tratamento de valores ausentes, para então apresentar as considerações finais.

Cap 4 - Descrição do problema: inicialmente apresenta-se uma breve fundamentação teórica acerca de problemas de otimização combinatorial, cobrindo sua definição, a identificação

e definição de problemas, a construção e solução de modelos, até a validação e implementação de soluções. Posteriormente, discutem-se trabalhos relacionados para posterior descrição da problemática em foco, indicando representações das soluções candidatas e também possíveis estratégias de inicialização, busca e avaliação das referidas soluções aplicáveis a este domínio. Por fim, apresentam-se as considerações finais.

Cap 5 - AGImp: neste capítulo são discutidos alguns trabalhos relacionados à imputação múltipla de dados no contexto de classificação de padrões. Em seguida é descrito o algoritmo genético para imputação de dados para otimizar classificadores baseados em aprendizado de máquina proposto, denominado aqui de AGImp. Também é apresentada a avaliação experimental, comparando a abordagem proposta com algoritmos de imputação de dados disponíveis. Por fim, apresenta-se um estudo da convergência e do impacto da parametrização da solução proposta neste domínio de aplicação, além das considerações finais.

Cap 6 - MOGAImp: este capítulo discute alguns trabalhos relacionados à imputação múltipla de dados e às medidas de desempenho mais recorrentemente utilizadas. Posteriormente, descreve-se a abordagem multiobjetivo proposta; a avaliação experimental conduzida; as comparações dos resultados obtidos pelos métodos de *baseline* com o método proposto; e as considerações finais.

Cap 7 - Extrapolações: neste capítulo são discutidas algumas lacunas na literatura e também são apresentadas extrapolações dos métodos propostos que vêm sendo desenvolvidos, como por exemplo a adaptação da abordagem multiobjetivo para o contexto de classificação multirrótulo; a utilização de uma estratégia evolucionária baseada em controle de parâmetros para evolução de soluções obtidas por métodos de imputação simples; e um método de imputação de dados voltado para análise de séries temporais baseado em programação genética, chamado aqui de GPImp. Por fim, são apresentadas as considerações finais.

Cap 8 - Conclusões: neste capítulo resume-se os desafios de pesquisa enfrentados neste projeto de doutorado, as contribuições técnico-científicas, as publicações advindas desta tese, bem como descrevem-se as restrições e potenciais trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONSIDERAÇÕES INICIAIS

É notório o crescimento acentuado do volume de dados produzidos nos últimos anos. Neste âmbito, faz-se necessário o desenvolvimento de técnicas para analisar os dados a fim de extrair conhecimento útil, conferindo vantagens competitivas às corporações (HAIR, 2014). No entanto, um problema recorrente é a incompletude das bases, com consequências danosas para o processo de análise de dados, haja vista que as técnicas desenvolvidas não são tolerantes à incidência de dados faltosos. Portanto, faz-se necessário a utilização de estratégias para lidar com esta problemática de forma a melhorar a qualidade do resultado final (NEWMAN, 2014).

Este capítulo apresentará uma breve fundamentação teórica sobre a análise de dados, enfatizando a problemática de dados faltosos e suas formas de tratamento. Por fim, é feita uma breve introdução aos modelos bioinspirados.

2.2 ANÁLISE DE DADOS

O processo de globalização da economia modificou as exigências do mercado, fazendo com que as empresas redirecionassem alguns dos seus investimentos do setor produtivo para o setor de serviços. A Tecnologia da Informação e Comunicação (TIC) é apontada por Margaria (2007) como a origem desta transformação. Foi este advento que possibilitou a derrubada de fronteira entre os países, o que fez emergir modelos de negócio como o *outsourcing*[1] (GROSSMAN; HELPMAN, 2005) e *crowdsourcing*[2] (HOWE, 2006; BRABHAM, 2008). Tais modelos tornaram o mercado mais competitivo, obrigando as empresas a buscarem por inovação dos serviços oferecidos (ARMELLINI; KAMINSKI; BEAUDRY, 2012).

Adicionalmente, o barateamento do *hardware* aumentou a capacidade de aquisição, armazenamento e processamento de dados; inundando pessoas e corporações com uma enxurrada de dados das mais variadas áreas do conhecimento como: economia, engenharia, sociologia, arqueologia, medicina e marketing (HAN; KAMBER, 2006). Atualmente fala-se em zettabytes – um bilhão de terabytes, fazendo surgir conceitos como o *big data*. Isto impõe desafios, não somente no que tange ao armazenamento e recuperação de informação, mas de efetivamente analisá-la à frente da concorrência.

Neste cenário, a análise inteligente dos dados traz ganhos consideráveis para a instituição/pessoa, pois possibilita o acesso à informação, que é o dado analisado e contextualizado

[1] Terceirização: subcontratação de outras empresas para a execução de determinada etapa do processo produtivo.
[2] Modelo produtivo on-line, distribuído e orientado à resolução de problemas, que utiliza o tempo livre de uma pessoa para direcionar a uma atividade construtiva.

(REZENDE et al., 2003). Além disso, permite a geração do conhecimento, o qual representa o resultado do processo de comparação e combinação de informações úteis e significativas. Dessa forma, diversas disciplinas propõem-se a analisar os dados de forma a obter conhecimento a partir deles, a exemplo da Análise Multivariada de Dados (FÁVERO et al., 2009) e da Extração de Conhecimento de Base de Dados, mais conhecido por *Knowledge-Discovery in Databases* (KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAN; KAMBER, 2006), ambas buscam a identificação de comportamentos e tendências nas mais diferentes áreas de conhecimento.

Segundo Pereira (2004), a análise multivariada é um vasto campo do conhecimento que envolve uma grande multiplicidade de conceitos estatísticos e matemáticos. Pode-se enxergar a análise multivariada de dados como sendo uma extensão das análises univariadas ou bivariadas, mas que estuda modelos em que todas as variáveis sejam aleatórias e inter-relacionadas, de modo que seus diferentes efeitos não possam ser interpretados separadamente (FÁVERO et al., 2009).

Os conhecimentos das análises uni, bi e multivariada são largamente utilizados no KDD, que agrega também outras áreas de conhecimento como: biologia, teoria da informação, economia; com o intuito de desenvolver métodos computacionais que permitam identificar, extrair, validar e utilizar conhecimentos úteis a partir dos dados disponíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O mesmo autor conceitua o KDD como um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de bases de dados. Rezende et al. (2003) elenca cinco etapas fundamentais para o processo de extração de conhecimento, como apresentado na Figura 1.

Para Rezende et al. (2003), KDD e Mineração de dados referem-se ao mesmo processo, compostos pelas seguintes etapas:

1. Identificação do problema;
2. Pré-processamento;
3. Extração de padrões;
4. Pós-processamento;
5. Utilização do conhecimento.

A primeira etapa envolve o entendimento do domínio de aplicação de forma a possibilitar a avaliação do processo como um todo por meio da validação do conhecimento extraído. A segunda etapa é de fundamental importância para o sucesso da extração de conhecimentos válidos e potencialmente úteis, ocupando cerca de 80% do tempo total do processo, haja vista

Figura 1 – Fluxo do processo de KDD.

Fonte: Rezende et al. (2003).

que os dados disponíveis para a análise não estão em um formato adequado para extração de conhecimento. O pré-processamento agrega métodos de tratamento, limpeza, transformação e redução do volume de dados, sendo que o tratamento de valores ausentes reside nesta etapa.

Em continuidade, aplicam-se os métodos de inteligência computacional para o reconhecimento de padrões. Han; Han e Kamber (2006) destaca os seguintes objetivos desta etapa:

- Classificação: prediz a qual classe um item pertence;
- Associação: identifica grupos de dados que apresentam coocorrência entre si;
- Agrupamento: mais conhecido por *clustering*, identifica grupos de dados associando-os aos rótulos;
- Regressão ou predição: mapeia valores dos dados em uma função preditiva, resultando em um ou mais valores reais.

Existem diversos métodos baseados em aprendizado de máquina desenvolvidos para satisfazer os objetivos listados acima. Algumas destas técnicas consistem na aplicação de um determinado algoritmo de extração de padrão, outras combinam diversos métodos visando prover uma melhor adaptabilidade e maior confiabilidade ao resultado final. Portanto, a etapa de processamento engloba a definição do objetivo e a escolha do algoritmo (WU et al., 2008).

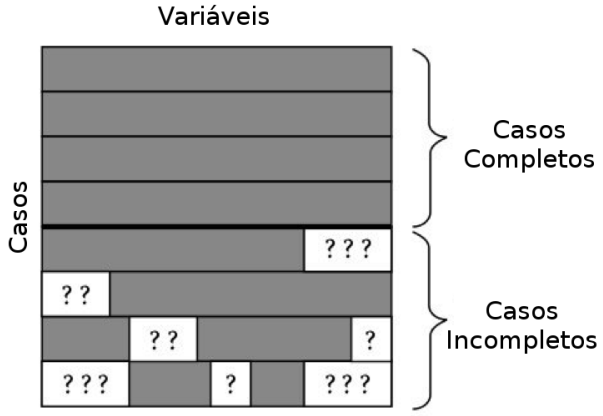
Posteriormente, há a avaliação e validação do conhecimento extraído, a fim de identificar possíveis falhas nas etapas anteriores (WITTEN; FRANK; HALL, 2011). Dessa forma,

garante-se a confiabilidade da quinta e última etapa, a de utilização do conhecimento, geralmente realizado por meio da incorporação dos resultados em um sistema de suporte à decisão. Como mencionado, a etapa de pré-processamento ocupa cerca de 80% do tempo de todo o processo, envolvendo o tratamento de valores ausentes, tema da próxima seção.

2.3 DADOS AUSENTES

Em um conjunto de dados, a ausência de itens em instâncias é denominado na língua inglesa de *missing data*; outros termos também são utilizados como *missing values* e *incomplete data* (LITTLE; RUBIN, 2002). Não há um consenso na tradução para a língua portuguesa, havendo diversos termos como dados faltosos, dados faltantes, dados incompletos, valores ausentes, dentre outros (OLIVEIRA, 2009; SILVA, 2010; VERONEZE, 2011b; FACELI et al., 2011); todos referenciando-se ao mesmo conceito, ilustrado na Figura 2.

Figura 2 – Exemplo de conjunto de dados composto por casos completos e casos incompletos.



Fonte: Adaptada de García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009).

Outra conceituação importante permite diferenciar casos completos e casos incompletos, sua definição é intuitiva, como mostra a Figura 2. Entende-se por casos completos instâncias que não possuem dados faltosos, enquanto casos incompletos, o contrário. Vale frisar que a ausência de valores é um problema recorrente no processo de análise de dados (HEERINGA; WEST; BERGLUND, 2010). Graham (2009) aponta que o aumento no interesse por este problema teve início em 1987 com a publicação do trabalho de Little e Rubin (1987); mesmo com a publicação de estudos importantes em um período anterior (DEMPSTER; LAIRD; RUBIN, 1977; HECKMAN, 1979; RUBIN, 1976).

Com ou sem valores ausentes, o objetivo da estatística é fazer, de forma eficiente, inferências válidas sobre uma população de interesse. No entanto, as técnicas de análise de dados não foram modeladas para serem tolerantes aos dados faltosos. Por este motivo, sua consequência é danosa para o processo, haja vista a imposição de um viés – tanto nas análises que descon-

sideram os dados com valores ausentes, quanto as que o tratam, como será apresentado adiante (GRAHAM, 2009; GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009). Quanto às causas, há uma ampla diversidade, sendo dependentes do mecanismo de aquisição de dados, Brown e Kros (2003) apresentam algumas categorias relacionadas às causas de VA:

- Fatores operacionais, tais como erros na entrada dos dados, estimativas, remoção acidental de campos de tabelas, entre outras;
- Recusa na resposta em pesquisas;
- Impossibilidade de aplicação de um determinado questionamento.

Os fatores operacionais são mais comuns no contexto de KDD. Exemplos representativos são: erros na entrada dos dados em sistemas de informação; problemas na etapa de *Datawarehousing* (*e.g.* ausência de determinados campos em uma das bases de dados que serão integradas para formar um *Data Warehouse*); falhas em uma rede de comunicação; e mau funcionamento do dispositivo de coleta de dados.

O segundo fator, recusa na resposta em pesquisas, afeta principalmente a análise de dados em ciências sociais e médicas. Como um exemplo para a segunda categoria, Allison (2001) aponta o questionamento da renda como um exemplo pertinente; para tal ele cita o conjunto de dados "*General Social Survey*" de 1994, com 2992 instâncias. Em 356 exemplos o entrevistado não respondeu os rendimentos. Da área médica podemos citar a omissão de determinado comportamento de risco como o consumo de drogas, dificultando o processo de diagnóstico.

A última categoria, impossibilidade de aplicação de um determinado questionamento, também acomete majoritariamente as áreas sociais e médicas. Por exemplo, o questionamento se o cliente deseja um plano de saúde com cobertura neonatal não se aplica a homens solteiros que não possuam dependentes. Definidos a problemática, suas possíveis causas e consequências, faz-se necessário a adoção de um formalismo para descrevê-lo. Esta proposta adotou a descrição encontrada em García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009) por ser a revisão mais atual e pela descrição estar em consonância com a literatura moderna de análise estatística com dados ausentes, conforme apresentado na subseção a seguir.

2.3.1 DEFINIÇÕES

No contexto da classificação de padrões com dados ausentes, uma instância ou exemplo é representado por um vetor de d atributos (contínuos ou discretos), como por exemplo, $x = [x_1, x_2, \ldots, x_i, \ldots, x_d]^T$ onde cada exemplo pertence a uma das c classes ou rótulos possíveis $C_1, C_2, ..., C_c$. Um conjunto de dados D é composto por N instâncias incompletas e rotuladas,

$$D = \{\mathbf{X}, \mathbf{T}, \mathbf{M}\} = \{(\mathbf{x}_n, \mathbf{t}_n, \mathbf{m}_n)\}_{n=1}^{N} \tag{2.1}$$

onde $x_n = [x_{1n}, x_{2n}, \ldots, x_{dn}]^T$ é o n-ésimo vetor composto por d atributos, rotulados como $t_n \in [C_1, C_2, \ldots, C_c]$; e $m_n = [m_{1n}, m_{2n}, \ldots, m_{dn}]^T$ indica quais atributos de entrada são desconhecidos em x_n. O vetor de indicação de dados faltosos, $\mathbf{m}_n$, é também denominado de "vetor de indicação de resposta"[3]. $\mathbf{X}$ é o conjunto dos dados de entrada, $\mathbf{M}$ é uma matriz binária que indica a ausência de valores; ambos têm dimensão $[1 \times N]$. De acordo com $\mathbf{M}$, $\mathbf{X}$ é divido em duas partes:

$$X = \{\mathbf{X}_0, \mathbf{X}_m\} \tag{2.2}$$

$\mathbf{X}_0$ e $\mathbf{X}_m$ são, respectivamente, os valores observados no conjunto de dados (completos), e as instâncias com valores ausentes. Tais definições fornecem o subsídio necessário para entender a relação entre os causadores dos dados faltosos e um efeito denominado pelos estatísticos de "Mecanismo de Ausência de Dados" [4] (LITTLE; RUBIN, 1987; LITTLE; RUBIN, 2002); os quais são descritos a seguir.

2.3.2 MECANISMOS DE AUSÊNCIA DE DADOS

A forma apropriada para tratar os valores ausentes depende, na maioria dos casos, em como os atributos tornaram-se ausentes. O mecanismo de ausência de dados tenta mapear isto e é caracterizado pela distribuição condicional de $\mathbf{M}$ dado $\mathbf{X}$:

$$p(\mathbf{M}|\mathbf{X}, \xi) = p(\mathbf{M}|\mathbf{X}_0, \mathbf{X}_m, \xi) \tag{2.3}$$

onde ξ denota o parâmetro desconhecido que define um dos três mecanismos de ausência de dados, a saber:

- Ausência completamente aleatória (*Missing completely at random* (MCAR)): situação que ocorre quando a probabilidade da variável ser faltosa é independente da própria variável ou por qualquer outra influência (valores ausentes ou observados) e pode ser expressa por:

$$p(\mathbf{M}|\mathbf{X}_0, \mathbf{X}_m, \xi) = p(\mathbf{M}|\xi) \tag{2.4}$$

  o que significa que a ausência da variável não depende dos valores de entrada pois, os exemplos disponíveis contém toda a informação para fazer inferências. Exemplos típicos do mecanismo MCAR são tubos de ensaio contendo uma amostra de sangue que quebram acidentalmente, logo, os parâmetros sanguíneos não podem ser mensurados. A razão para a ausência de dados é completamente aleatória – a probabilidade que uma observação seja ausente não é relacionada a qualquer outra característica do indivíduo.

[3] Tradução de "response indicator vector"

[4] Tradução literal de "*Mechanisms of Missingness*".

- Ausência aleatória (*Missing at random* (MAR)): a ausência de dados é independente dos valores ausentes, mas o padrão de ausência é predita por outras variáveis da base de dados. A condição para ser considerada MAR é expressa pela relação:

$$p\left(\mathbf{M}|\mathbf{X}_0,\mathbf{X}_m,\xi\right) = p\left(\mathbf{M}|\mathbf{X}_0,\xi\right) \tag{2.5}$$

  a ausência da variável depende apenas de valores observados nos dados de entrada (casos completos). Um exemplo é a falha ocasional de um sensor devido a uma queda de energia, interrompendo o processo de aquisição. Neste exemplo, as variáveis atuais onde os dados estão faltando não são os causadores da incompletude, pois a causa da ausência está em uma influência externa.

- Ausência não aleatória (*Not Missing at Random* (NMAR)): o padrão de dados faltosos não é aleatório e depende do próprio valor ausente, a qual pode ser descrita por meio da equação:

$$p\left(\mathbf{M}|\mathbf{X}_0,\mathbf{X}_m,\xi\right) \neq p\left(\mathbf{M}|\mathbf{X}_0,\xi\right) \tag{2.6}$$

  em contraste com o padrão MAR, a variável ausente no caso MNAR não pode ser predita apenas levando-se em consideração as variáveis do conjunto de dados. Por exemplo, se um sensor não consegue adquirir informação fora de uma determinada faixa, este dado é faltoso devido ao MNAR. Então, diz-se que os dados foram censurados. Portanto, informações importantes são perdidas, e não há nenhum método para lidar corretamente com este tipo de falta. Outro exemplo que se faz interessante notar é quando um atributo x_{nb} é computado a partir de um outro atributo, x_{na} que está ausente; logo, x_{nb} também estará ausente e não haverá, no conjunto de dados, informação que leve à inferência do seu valor.

Para Schafer e Graham (2002), há ainda um quarto mecanismo que recai no exemplo apresentado para o padrão MAR, o de valores fora de uma determinada faixa. Contudo, a presente proposta irá ater-se somente às três categorias acima descritas, conforme padrão encontrado na literatura. Acerca dos padrões MCAR ou MAR, é um consenso denominá-los de padrões ignoráveis. Este fato é importante pois, quando ele ocorre, os pesquisadores podem ser indiferentes quanto à natureza dos dados faltantes. Em outras palavras, tais mecanismos são fáceis de manipular, visto que seus efeitos nos modelos estatísticos, e atualmente, nos de Aprendizado de Máquina, estão disponíveis para os analistas (MCKNIGHT et al., 2007; GRAHAM, 2009). Ainda na análise dos padrões MCAR e MAR, a simples comparação entre as Equações 2.4 e 2.5 torna possível atestar que o MCAR possui menos parâmetros, logo a estimação é mais simples do que no mecanismo MAR. O qual possui um modelo que descreve a ausência dos dados a partir das informações contidas em $\mathbf{X}$ (Eq. 2.2).

Em contrapartida, o mecanismo NMAR é dito *não-ignorável*, neste caso não há informação no conjunto de dados que permita a modelagem do comportamento do mecanismo de ausência. Consequentemente, o efeito deste padrão na construção do modelo, seja ele estatístico ou de aprendizado de máquina, é difícil de se estimar. Portanto, conhecer o mecanismo de ausência de dados auxilia o analista no entendimento da natureza dos dados faltosos e o respectivo impacto nas análises subsequentes.

A categorização de um atributo em qual mecanismo de ausência de dados se enquadra é realizada por exclusão. Primeiro avalia-se se o mecanismo é MCAR, caso os requisitos não sejam atendidos, testa-se o MAR, e por exclusão o NMAR. A avaliação do mecanismo MCAR dá-se pelos métodos propostos por Little (1988) e Chen e Little (1999), enquanto para a avaliação dos demais métodos não há um método formal dispostos na literatura. Para fins práticos, a maior parte das pesquisas envolvendo tratamento de valores ausentes assume que os dados faltosos são regidos pelo mecanismo MAR ou MCAR. Mcknight et al. (2007) apresentam um quadro que resume as características intrínsecas dos mecanismos de ausência dos dados e as possíveis causas em pesquisas com aplicação de questionários (Quadro 2.1).

Quadro 2.1 – Quadro comparativo entre as possíveis causas dos mecanismos de ausência de dados em pesquisas com aplicação de questionários.

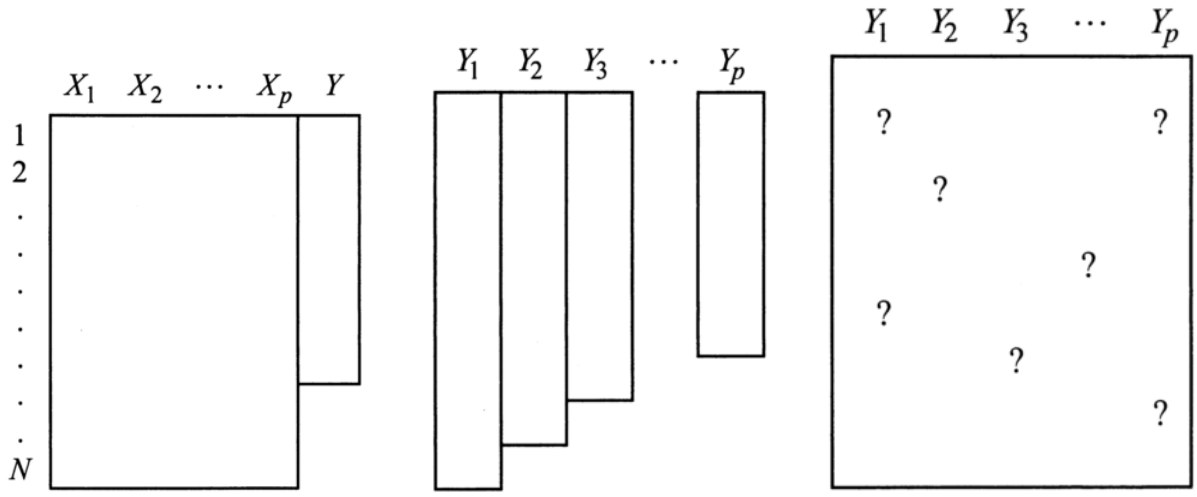
Mecanismo / **Situação**	**MCAR**	**MAR**	**MNAR**
Variável (Item)	Indivíduos omitem respostas aleatoriamente.	Indivíduos omitem respostas que podem ser conseguidas por outras respostas.	Indivíduos não respondem itens indiscriminadamente.
Indivíduos	Faltam dados de indivíduos aleatoriamente.	Faltam dados de indivíduos, mas que são relacionados com os dados demográficos disponíveis.	Faltam dados de indivíduos e são relacionados com os dados demográficos não medidos.
Ocasiões	Indivíduos aleatoriamente não se apresentam na sessão.	Indivíduos que se desempenham mal na sessão anterior e não se apresentam na sessão seguinte.	Indivíduos que estão se desempenhando mal na sessão atual e deixam de participar.

Fonte: Adaptada de Mcknight et al. (2007).

Este quadro comparativo possibilita uma melhor compreensão dos mecanismos de ausência de dados. Contudo, esta não é a única categorização, há ainda a determinação do padrão de dados faltosos, como apresentado a seguir.

2.3.3 PADRÃO E QUANTIFICAÇÃO DE DADOS FALTOSOS

Os dados ausentes podem ser caracterizados em uma série de padrões, que identificam se há ou não um comportamento comum quanto à forma como os dados foram observados (MCKNIGHT et al., 2007). Os principais padrões de ausência de dados discutidos por Schafer e Graham (2002) são apresentados na Figura 3.



(a) padrão univariado (b) padrão monotônico (c) padrão arbitrário

Figura 3 – Padrão de dados faltosos em um conjunto de dados retangular.

Fonte: Schafer e Graham (2002).

Em cada caso, linhas correspondem aos exemplos e colunas, aos atributos. O padrão univariado descreve o caso em que as instâncias possuem apenas um atributo ausente, como mostra a Figura 3a. Um exemplo deste padrão é a negligência de um determinado item de um questionário, como a informação da renda mensal.

No caso do padrão monotônico, os dados passam a faltar a partir de um determinado ponto e, além disso, os exemplos com dados faltosos seguem um padrão particular – como pode ser notado na Figura 3b. Este comportamento é comum em estudos longitudinais[5].

O terceiro e último padrão é dito arbitrário ou geral, quando atributos são negligenciados de forma aleatória no conjunto de exemplos. Outra informação pertinente para a escolha da abordagem de tratamento de dados faltosos é a quantificação de VA. Para Mcknight et al. (2007) é possível realizar cinco observações levando-se em conta o número de:

[5] Metodologia de pesquisa que observa determinado número de variáveis de forma periódica em um determinado período de tempo.

1. Atributos com dados ausentes;
2. Instâncias com dados ausentes;
3. Valores ausentes em um atributo específico;
4. Valores ausentes em um conjunto de atributos específicos;
5. Valores ausentes em todo o conjunto de dados.

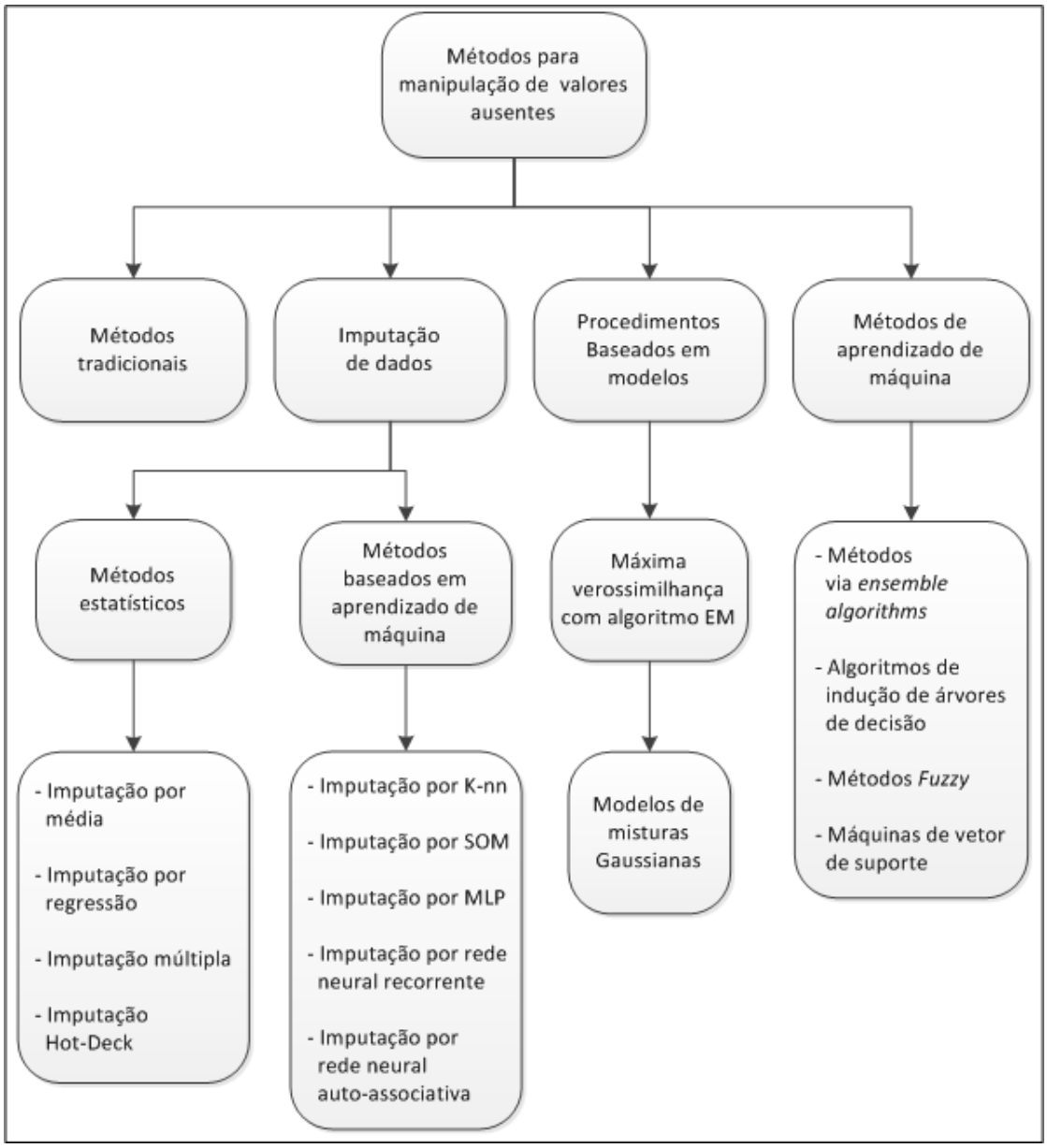
Estas informações auxiliam na escolha da técnica de tratamento de valores ausentes, como será melhor apresentado na seção a seguir.

2.3.4 TRATAMENTO DE VALORES AUSENTES

Há várias formas para lidar com valores ausentes, com um número considerado de técnicas já propostas na literatura, no entanto, poucas ganharam destaque. De forma geral, é possível dividir os métodos de tratamento de valores em quatro classes conforme apresentado no desenho esquemático disposto na Figura 4, tais métodos encontram-se descritos a seguir.

- **Abordagens tradicionais:** também chamadas de *análise de caso completo*, tratam os dados ausentes por meio da simples omissão, seja de instâncias ou atributos, que contenham valores ausentes. São exemplos notórios: *listwise deletion* e *pairwise deletion*;
- **Imputação:** substituem o valor associado ao dado faltoso, normalmente *null* ou "?", por um valor plausível. Subdividem-se em métodos estatísticos e métodos baseados em aprendizado de máquina. Os primeiros utilizam medidas estatísticas para a estimação do(s) valor(es) a ser(em) imputado(s), enquanto a segunda classe utiliza-se de algoritmos de aprendizado de máquina para predição destes valores. Alguns exemplos são: imputação simples, imputação múltipla, imputação estimada por vizinhança etc;
- **Modelos:** métodos iterativos que visam utilizar técnicas para estimar a máxima verossimilhança de forma a inferir a função de distribuição conjunta de cada atributo, para auxiliar na estimativa do valor a ser imputado. Apesar de realizar a imputação, alguns autores não incluem este método na classe de imputação. São exemplos de métodos desta categoria: *Expectation-Maximization algorithm*, *Gaussian Mixture Models*;
- **Métodos de aprendizado de máquina:** evadem-se da imputação explícita, tendo como alvo o desenvolvimento/adaptação dos algoritmos de aprendizado de máquina para aumentar sua robustez quando à incidência de dados faltosos. Os métodos baseados em combinação de classificadores (*ensemble classifiers*) e métodos *fuzzy* ilustram algumas técnicas desta categoria.

Figura 4 – Diagrama esquemático com o resumo dos principais métodos para classificação de padrões com valores ausentes.



Fonte: Adaptada de García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009).

Zhang (2010) também propõem uma categorização dos métodos de imputação baseado na quantidade de imputações, conforme apresentado a seguir:

- **Imputação simples**: fornece uma única estimativa para cada valor ausente. Pertencem a esta categoria a imputação por média, o tratamento interno do algoritmo C4.5, *k-Nearest Neighbor Imputation* (KNNI) *etc*;
- **Imputação Múltipla**: estima possíveis valores para imputação baseando-se em medidas apropriadas para verificação da precisão a fim combinar estas estimativas ao valor final, o

método de imputação múltipla proposto por Rubin (1987) é o exemplo mais ilustre desta categoria;

- **Imputação fracionada**: representa um meio termo entre as duas primeiras categorias, proposto inicialmente por Kang, Koehler e Larsen (2007) e com o representante mais conhecido a imputação funcional paramétrica (KIM, 2011);

- **Imputação iterativa**: basicamente, utiliza o mecanismo de geração-e-teste levando em consideração informações úteis (incluindo os casos incompletos). Variantes do KNNI que incluem processos iterativos e métodos baseados em computação bioinspirada pertencem a esta categoria.

Seja qual for a abordagem utilizada, o objetivo é diminuir o viés imposto pelos dados faltosos que inerentemente afeta o resultado da análise de dados, haja vista, como mencionado anteriormente, que as técnicas de análise de dados não foram modeladas para lidar diretamente com VA.

2.4 COMPUTAÇÃO EVOLUCIONÁRIA

Na computação, pesquisadores utilizam ideias extraídas da observação da natureza para desenvolver soluções baseadas em sistemas computacionais desde a invenção do computador (RUSSELL; NORVIG, 2009). Nas décadas de 70 e 80 percebeu-se uma tendência em desenvolver diferentes algoritmos que implementam estratégias bioinspiradas (FOGEL, 1999; SCHWEFEL, 1981; HOLLAND, 1992) e atualmente investiga-se também a utilização de materiais naturais, como átomos e estruturas de DNA [6], para realizar a computação (NIELSEN; CHUANG, 2011).

Tais abordagens estão contidas no conceito de Computação Natural, a qual pode ser definida como a versão computacional do processo de extração de ideias da natureza para desenvolver sistemas computacionais (De Castro, 2007). O mesmo autor define este campo de estudo em três ramificações:

- Computação bioinspirada: faz uso da natureza como forma de inspiração para o desenvolvimento de técnicas de resolução de problemas. Sua ideia principal consiste na observação da natureza com o objetivo de extrair padrões e comportamentos e basear-se neles para resolver problemas complexos a fim de desenvolver ferramentas computacionais ou algoritmos;

- Simulação e emulação da natureza por meio da computação: seus produtos podem ser usados para simular vários fenômenos naturais, aumentando assim a compreensão da natureza e as percepções sobre modelos computacionais;

[6] Acrônico na língua inglesa de ácido desoxirribonucleico.

- Computação com materiais naturais: constituem um novo paradigma de computação que surge para substituir ou complementar os computadores atuais à base de silício.

A primeira ramificação é de longe a mais trabalhada por sua larga aplicabilidade nos mais diversos domínios, e também pelo fato dos modelos tradicionais não conseguirem obter uma resolução satisfatória para um determinado problema. Dentre as abordagens mais conhecidas estão a computação evolucionária e a inteligência de enxame. Tais abordagens têm sido aplicadas com sucesso em uma ampla gama de tarefas computacionais em otimização, design e modelagem de sistemas, e também têm se mostrado como métodos eficientes para extração de padrões.

Atualmente, a família de Algoritmos Evolucionários (AE) incluem alguns membros históricos: algoritmos genéticos, programação evolucionária, programação genética, evolução diferencial e otimização por enxame de partículas (GOLDBERG, 1989; BÄCK, 1996; BANZHAF et al., 1998; KENNEDY; EBERHART; SHI, 2001; PRICE; STORN; LAMPINEN, 2005). Eles diferem em alguns detalhes técnicos, terminologias ou na fonte de inspiração, mas eles possuem alguns itens em comum, como por exemplo, a população deve: passar informação gênica à prole, apresentar variabilidade genética e passar pela seleção natural (De Castro, 2007).

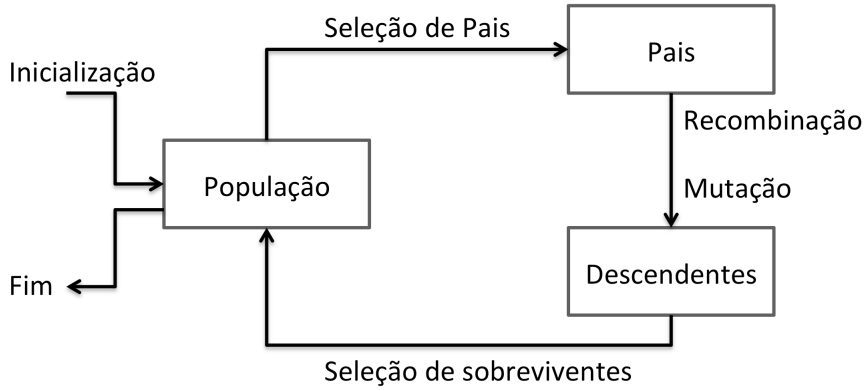
Por população entende-se um grupo de indivíduos, o qual é formado por uma tupla, genótipo e fenótipo, no qual o primeiro item diz respeito às informações genéticas que, dentre outras coisas, proverá a hereditariedade; enquanto a segunda representa a manifestação da característica física do organismo (HOLLAND, 1992).

Neste ponto, convém observar que problemas de otimização combinatorial estão relacionados com a alocação eficiente de recursos limitados para atender objetivos desejados. As variáveis de decisão podem ser contínuas ou discretas e podem ser limitadas por restrições sobre as mesmas, limitando o número de alternativas possíveis a serem consideradas factíveis (ROTHLAUF, 2011).

Um indivíduo de um algoritmo evolucionário representa uma solução para o problema a ser resolvido, nele estão codificadas as variáveis de decisão, a qual está associada à uma ou mais função(ões) objetivo, aqui chamada de função de aptidão ou *fitness*, que indica o quão ele é apto à se desenvolver sob determinadas condições.

Eiben e Smith (2015) destacam que algoritmos evolucionários são facilmente adaptáveis de uma aplicação para outra pois apenas dois componentes são dependentes, a forma pela qual o genótipo é convertida no fenótipo e a função de aptidão. Sendo assim, o primeiro passo do projeto de um algoritmo evolucionário é definir a representação do indivíduo, onde escolhe-se a estrutura de dados apropriada; o segundo passo é definir a função de aptidão levando-se em consideração requisitos específicos do problema. O último passo é definir os operadores que caracterizam o processo evolucionário. A Figura 5 sintetiza o funcionamento básico de um algoritmo evolucionário, destacando os operadores utilizados.

Figura 5 – Esquema que representa o funcionamento básico de um algoritmo evolucionário.



Fonte: Adaptada de Eiben e Smith (2003).

A primeira etapa é a **inicialização** das soluções candidatas, aqui representadas pelos indivíduos por meio de seu cromossomo. Geralmente utiliza-se uma inicialização pseudo-aleatória dos genes, respeitando os limites inferior e superior das variáveis. Com a população inicializada, indivíduos que irão se tornar pais **selecionados**, geralmente levando-se em consideração sua função de aptidão pois infere-se que os indivíduos mais aptos ao meio são os que possuem maior probabilidade de gerar descendentes.

Os indivíduos selecionados então trocam material genético para gerar a prole, aqui representado pelo processo de **recombinação**. Convém destacar que durante o processo de evolução, o indivíduo pode sofrer alterações no seu material genético, fenômeno denominado de **mutação**, que visa prover variabilidade às soluções fornecendo novos pontos de busca para o problema.

Com a população de descendentes completa, **selecionam-se os sobreviventes** para compor a população da próxima geração. Este processo continua até atingir um dos critérios de parada, como por exemplo: número máximo de gerações, número máximo de consultas à função de aptidão, erro mínimo. Como os operadores de seleção são estocásticos, é possível que ocorra a perda de soluções ótimas ou próximos dos pontos ótimos; para que isso não ocorra, é possível aplicar o operador denominado de **elitismo**, que passa automaticamente os melhores indivíduos para a próxima geração.

Os operadores acima descritos conferem aos algoritmos genéticos duas características básicas, conhecidas por *exploration* e *exploitation* – que definem a exploração do espaço de busca como um todo e a exploração de locais próximos ao ótimo, respectivamente (EIBEN; SCHIPPERS, 1998).

Por este motivo, aliado ao fato de sua fácil implementação e adaptação a diversos nichos de aplicação, os algoritmos evolucionários vêm sendo largamente adotados em problemas de busca e otimização. As subseções a seguir examinam dois algoritmos evolucionários, o algo-

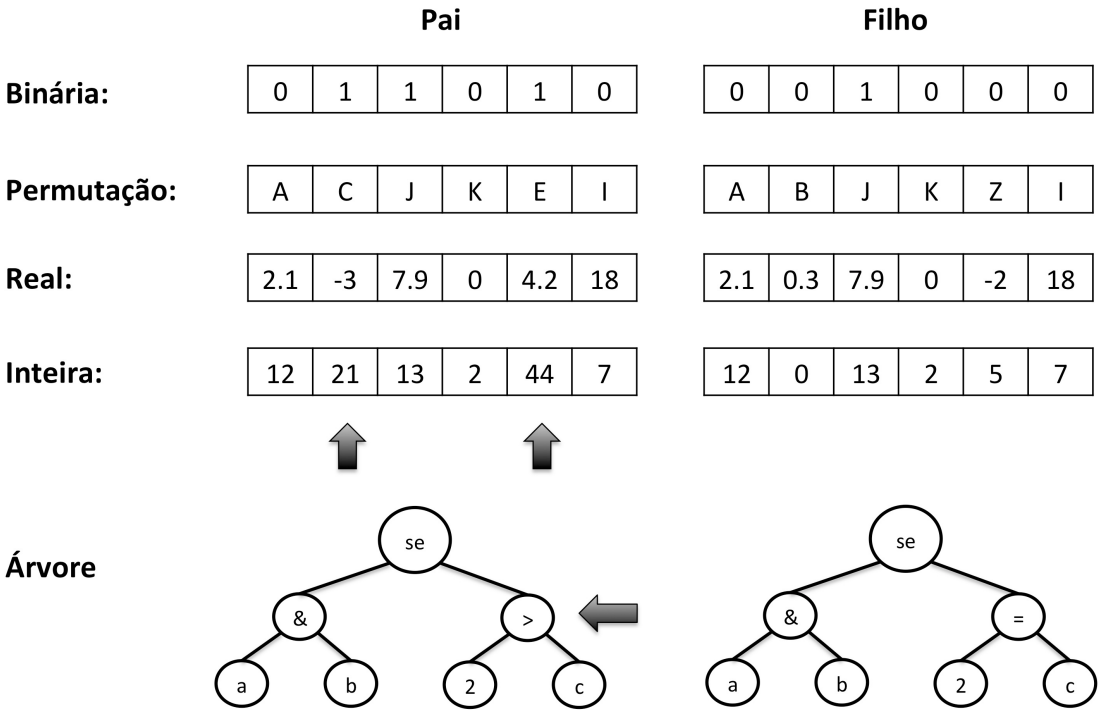
ritmo genético e a programação genética.

2.4.1 ALGORITMOS GENÉTICOS

Um dos maiores representantes da computação evolucionária são os algoritmos genéticos (De Castro, 2007). Como apresentado anteriormente, o projeto de algoritmos evolucionários envolvem dois aspectos: i) determinar a codificação do indivíduo e a função de aptidão; ii) definir os parâmetros do algoritmo.

A codificação do indivíduo representa a forma pela qual o genótipo será mapeado no fenótipo correspondente e as estruturas de dados envolvidas no processo. A Figura 6 apresenta algumas codificações armazenadas em vetor ou em árvore.

Figura 6 – Representação de soluções como genótipos e exemplificação de mutação.



Fonte: Adaptada de Eiben e Smith (2015).

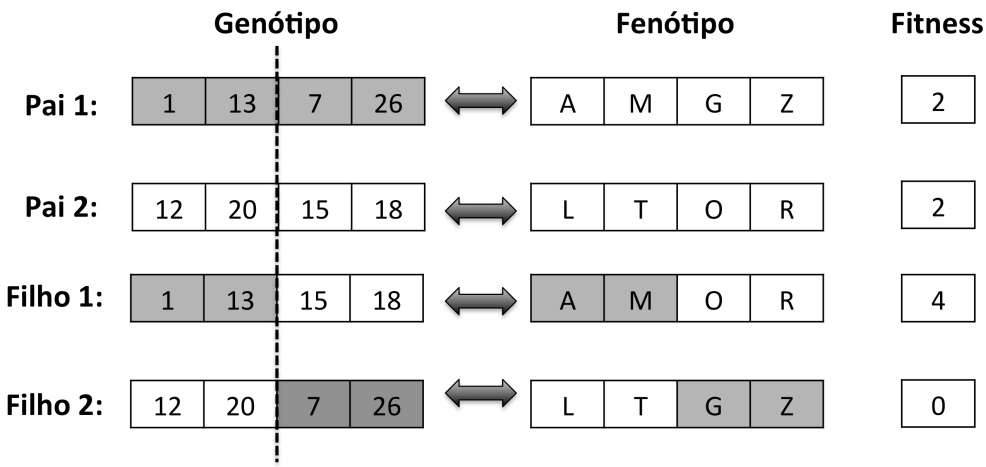
No exemplo de codificações apresentado pela Figura 6, têm-se cinco esquemas usuais para o genótipo: binária, permutação, real, inteira e em árvore. Uma das estruturas de dados mais usadas para armazenamento é o vetor, onde seu tamanho é determinado pelo número de variáveis de interesse e a precisão requerida pelo problema. Convém ressaltar que a escolha da codificação e da estrutura de dados utilizada tem impacto significativo na seleção dos operadores, uma vez que há operadores específicos para cada esquema de codificação.

Ainda na Figura 6, as setas apresentam pontos onde o operador de mutação é aplicado, consequentemente o valor destes genes modificam-se, contudo, o mecanismo que implementa este operador varia de acordo com o esquema de codificação. No tocante à representação do

indivíduo, o fenótipo pode ser idêntico ao genótipo, sobretudo quando utiliza-se a codificação real ou inteira.

No entanto, dada a flexibilidade inerente a esta classe de algoritmos, alguns problemas podem utilizar-se de outro esquema de codificação a fim de reaproveitar operadores. Por exemplo, a Figura 7 apresenta um exemplo onde a codificação do genótipo difere do fenótipo, requerendo um processo de mapeamento.

Figura 7 – Exemplo de mapeamento genótipo-fenótipo e de recombinação.



Fonte: Elaborada pelo autor.

No exemplo apresentado na Figura 7 o objetivo é encontrar a palavra "AMOR", para tal, os indivíduos têm seu genótipo codificado em inteiros de 1 à 26, onde cada inteiro representa uma letra ($A = 1, B = 2, \ldots Z = 26$) e o mapeamento do genótipo em fenótipo dá-se pela consulta na tabela. Neste tipo de aplicação, a função de aptidão pode ser obtida por meio das métricas de similaridade entre cadeias de caracteres, no exemplo dado, a cada acerto de um gene, o *fitness* é acrescido de uma unidade.

Outro conceito apresentado na Figura 7 é o de cruzamento, onde os indivíduos trocam material genético, gerando-se os descendentes. De posse dos conceitos pertinentes à codificação dos indivíduos, função de aptidão e uma visão geral dos operadores de mutação e recombinação, a subseção a seguir apresentará o funcionamento do algoritmo genético dito canônico e os parâmetros necessários.

2.4.1.1 FUNCIONAMENTO

O algoritmo genético canônico incorpora o fluxo apresentado na Figura 5 e é descrito no Algoritmo 1, o qual tem como entrada os parâmetros qualitativos e quantitativos de um algoritmo genético. Os qualitativos referem-se aos valores simbólicos como os operadores de seleção, cruzamento e mutação, os quais têm um domínio finito sem uma métrica de distância ou ordenamento entre eles, *e.g.,* $\{onepoint, uniform, n-point\}$; e os quantitativos são majo-

ritariamente numéricos, *e.g.,* taxa de cruzamento, taxa de mutação, tamanho da população *etc*. Ainda nessa seção, os parâmetros quantitativos e qualitativos são apresentados e seu impacto no desempenho do AG são discutidos.

Algoritmo 1: Algoritmo genético canônico

Entrada: *Param_Set*
Saída : *Sol_Set*

```
1  Inicializa a população inicial;
2  enquanto critério_parada = falso faça
3      Avaliar a função de aptidão;
4      enquanto tam_desc ≠ n_individuos faça
5          Selecionar indivíduos para cruzamento;
6          Realizar o cruzamento entre os indivíduos selecionados;
7          Aplicar mutação;
8          Adicionar indivíduos na população de descendentes;
9      fim
10     Reorganizar populações;
11 fim
```

Em prosseguimento à análise do Algoritmo 1, a saída é composta por um conjunto de soluções representada pela última população sobrevivente. Definida a entrada e a saída, o processo de evolução das soluções começa pela inicialização da população de indivíduos, onde atribui-se os valores aos genes dos indivíduos, geralmente de forma aleatória. Com todos os indivíduos instanciados, avalia-se o critério de parada (Linha 2), caso ele não tenha sido alcançado, gera-se uma nova população (Linhas 4-9).

Nesta etapa, primeiro selecionam-se os indivíduos para o cruzamento (Linha 5); realiza-se a recombinação entre os indivíduos selecionados (Linha 6); aplica-se o operador de mutação, obedecendo uma probabilidade de ocorrência (Linha 7); e adicionam-se os indivíduos gerados na população de descendentes. Este processo é realizado até a população de descendentes estar completa, então o algoritmo reorganiza a população (Linha 10), onde os descendentes passam a ser a população de pais. Como visto, o Algoritmo 1 possui um conjunto de parâmetros como entrada. A subseção a seguir apresenta conceitos pertinentes sobre parametrização de algoritmos evolucionários, classe a qual os algoritmos genéticos pertencem.

2.4.1.2 PARAMETRIZAÇÃO

A diferenciação entre parâmetros quantitativos e qualitativos apresentados na subseção anterior auxilia na diferenciação entre os conceitos de Algoritmo Genético e Instância de um Algoritmo Genético. Esta visão é baseada considerando-se os parâmetros qualitativos como alto-nível, os quais definem a estrutura principal de um algoritmo evolucionário; e os parâmetros quantitativos como baixo-nível, que definem variações de um algoritmo (EIBEN; SMIT, 2011).

Baseando-se nessa convenção, um **algoritmo evolucionário** possui seus parâmetros qualitativos instanciados, mas não possui valores atribuídos aos parâmetros quantitativos. Quando todos os parâmetros estão especificados, tem-se então uma **instância de um algoritmo evolucionário**.

O Quadro 2.2 exemplifica estes conceitos, nele são mostrados três exemplos, os algoritmos genéticos A1 e A2 são ditos idênticos uma vez que possuem os mesmos parâmetros qualitativos, mas são instanciados diferentemente devido possuírem diferentes valores para os parâmetros quantitativos, analogamente, o algoritmo A3 difere de A1 e A2 pois possui atributos diferentes para os parâmetros qualitativos.

Quadro 2.2 – Exemplo de especificação de parâmetros.

	A_1	A_2	A_3
Parâmetros Qualitativos			
Representação	Binária	Binária	Real
Recombinação	1-ponto	1-ponto	Média
Mutação	Bit-flip	Bit-flip	Gaussiana $N(0,\sigma)$
Seleção de Parentes	Torneio	Torneio	Aleatório Uniforme
Seleção de sobreviventes	Geracional	Geracional	(μ,λ)
Parâmetros Quantitativos			
ρ_m	0.01	0.1	0.05
σ	n.a.	n.a.	0.1
ρ_c	0.5	0.7	0.7
μ	100	100	10
λ	n.a	n.a	70
κ	2	4	n.a.

Fonte: Adaptada de Eiben e Smit (2011).

O Quadro 2.2 lista os atributos de um algoritmo evolucionário, seus parâmetros qualitativos foram brevemente apresentados na subseção anterior e podem ser resumidos da seguinte forma:

- Representação: refere-se ao tipo de codificação do genótipo, seja ela binária, real, inteira *etc*;
- Recombinação: também chamada de operador de cruzamento, especifica a lógica utilizada para recombinar os genes dos pais a fim de se gerar os descendentes;
- Mutação: operador responsável por modificar a informação gênica de acordo com uma probabilidade de ocorrência;
- Seleção de parentes: refere-se ao mecanismo utilizado para selecionar indivíduos para a recombinação;

- Seleção de sobreviventes: especifica a lógica utilizada para selecionar os indivíduos aptos para passar para a próxima geração.

Os parâmetros quantitativos comuns a todos os algoritmos evolucionários são:

- Taxa de mutação (p_m): concerne à probabilidade de aplicação do operador de mutação, altas taxas de mutação torna o processo de busca aleatório;
- Taxa de cruzamento (p_c): denota a taxa de indivíduos a serem submetidos à recombinação;
- Tamanho da população (μ): refere-se ao número de indivíduos que perfazem uma população;

Outros parâmetros podem ser necessários, dependendo dos parâmetros qualitativos utilizados, como por exemplo o tamanho do passo de mutação (σ) é requerido por alguns operadores de mutação, tal como a mutação Gaussiana; o tamanho da prole (λ), que especifica o número de indivíduos descendentes é necessário dependendo do operador de seleção de parentes; e o número de indivíduos utilizados no torneio (κ) que é apenas necessário quando usa-se o operador de torneio para seleção de parentes. A escolha dos valores influencia fortemente no desempenho e é considerada uma tarefa não trivial, por conseguinte, o processo de parametrização é um dos principais desafios para os projetistas de AE. Neste contexto, duas abordagens para escolher os valores para parâmetros são propostas por Eiben, Hinterding e Michalewicz (1999):

- Sintonia de parâmetros: ocorre quando (bons) valores para os parâmetros são estabelecidos antes da execução do algoritmo e eles permanecem inalterados durante a execução.
- Controle de parâmetros: ocorre quando os (bons) valores para os parâmetros são estabelecidos durante a execução de um algoritmo evolucionário, onde valores iniciais são fornecidos no começo da execução e então eles sofrem mudança no decorrer do processamento.

Estas duas abordagens são claramente relacionadas, pois visam obter valores ótimos (ou sub-ótimos) para os parâmetros, mas elas possuem diferenças e indicações. De forma usual, diz-se que a sintonia de parâmetro é algo obrigatório de ser realizado, enquanto o controle de parâmetro é algo desejável, isso é dito pois um AE pode ser executado sem que ocorram mudanças nos valores dos seus parâmetros, mas não pode ser executado sem que todos os parâmetros estejam definidos previamente (KARAFOTIAS; HOOGENDOORN; EIBEN, 2015).

A escolha de qual abordagem será utilizada para a parametrização depende de vários fatores, como por exemplo: *expertise* do projetista, natureza do problema (controle de parâmetro é indicado para problemas dinâmicos, por exemplo), conhecimento *a priori* do problema

etc. Ambas as abordagens necessitam de uma medida de desempenho para avaliar a parametrização, dado um conjunto de problemas-teste. Conforme a terminologia em consenso, o termo **utilidade** é adotado para denotar a qualidade do vetor de parâmetros, para tal, é necessário medir o desempenho do AE. Isto é feito levando-se em conta a qualidade das soluções e a velocidade do algoritmo. Diversas medidas que podem correlacionar estes dois itens encontram-se dispostos na literatura, Eiben e Jelasity (2002) discutem os prós e contras de diversas medidas de tempo de execução.

Basicamente, devido à natureza estocástica dos AE, várias execuções sobre um mesmo problema são necessárias para se realizar uma boa estimativa de performance. Agregando tais informações sobre um determinado número de execuções, algumas medidas de desempenho comumente usadas para avaliar algoritmos evolucionários são obtidas, como por exemplo: taxa de sucesso; número médio de avaliações da função de aptidão ou média da melhor aptidão (EIBEN; SMITH, 2003).

O estudo da parametrização não é restrito a se obter uma vetor de parâmetros ótimo para o problema em questão, mas está intrínsecamente ligado a uma melhor compreensão do comportamento do algoritmo por meio de experimentos controlados, onde estuda-se os efeitos das peculiaridades do problema e as características do algoritmo (incluindo seus parâmetros) sobre o seu comportamento durante a execução do método (EIBEN; SMIT, 2011). Diversos métodos para sintonização e controle de parâmetros têm sido propostos e analisados pela literatura. Em Eiben e Smit (2011) e Karafotias, Hoogendoorn e Eiben (2015) encontram-se a descrição de alguns métodos, bem como algumas tendências e desafios encontrados na área.

Como visto, diversos fatores tornam o projeto de um algoritmo genético uma tarefa não trivial. É necessário escolher a codificação do indivíduo e todos os operadores e variáveis que definem o seu comportamento de forma a prover o melhor desempenho para se otimizar uma função objetivo, contudo, diversos problemas de otimização reais envolvem vários objetivos ao mesmo tempo, um exemplo clássico é a compra de um computador, onde deseja-se minimizar o investimento e maximizar o desempenho da máquina a ser adquirida. Algoritmos evolucionários, sobretudo os algoritmos genéticos, vêm sendo aplicados com sucesso nesta gama de problemas, conforme apresentado na subseção a seguir.

2.4.1.3 ABORDAGENS MULTIOBJETIVOS

Formulações multiobjetivos, também conhecidas por otimização multi-critério, multi-desempenho ou problema de otimização de vetores[7], são modelos realísticos para vários problemas reais de otimização. Em diversos cenários, os objetivos que estão sendo considerados são conflitantes entre si, pois quando otimiza-se uma solução em relação à um objetivo específico, o resultado pode ser inaceitável para outros objetivos. Matematicamente, Problemas de

[7] Tradução do autor de “vector optimization problem”

Otimização Multiobjetivo (POM) podem ser definidos como (ZHOU et al., 2011):

$$\begin{aligned} &minimize \;\; F(x) = (f_1(x), \ldots f_m(x))^T \\ &s.a. \\ &x \in \Omega, \end{aligned} \tag{2.7}$$

onde Ω é o espaço de busca das variáveis de decisão e $x \in \Omega$ é o vetor com as variáveis de decisão. $F(x)$ consiste nas m funções objetivos $f_i : \Omega \to R, i = 1, \ldots, m$, onde R^m é o espaço das funções objetivo. Como dito, os objetivos dispostos na Eq. 2.7 podem ser conflitantes entre si. Para Konak, Coit e Smith (2006), há duas abordagens gerais para lidar com POM, a primeira consiste em combinar as funções em uma única função objetivo ou tratá-las separadamente, movendo as outras para o conjunto de restrições. A segunda abordagem geral utiliza o conceito de otimalidade de pareto, a qual pode ser definida como segue (MIETTINEN, 1999; DEB; KALYANMOY, 2001).

Definição 1. *Diz-se que um vetor* $u = (u_1, \ldots, u_m)^T$ *domina outro vetor* $v = (v_1, \ldots, v_m)^T$*, denotado por* $u \prec v$*, se* $\forall\, i \in \{1, \ldots, m\}$, $u_i \leq v_i$ *e* $u \neq v$.

Definição 2. *Uma solução factível* $x^* \in \Omega$ *de um problema (Eq. 2.7) é chamada de solução ótima de Pareto, se* $\nexists\, y \in \Omega$ *tal que* $F(y) \prec F(x^*)$*. O conjunto de todas as soluções ótimas de Pareto é chamado de Conjunto de Pareto (CP), expresso por*

$$CP = \{x \in \Omega | \nexists\, y \in \Omega, F(y) \prec F(x^*)\}.$$

A imagem do CP no espaço das funções objetivo é chamada de Fronteira de Pareto (FP), definida por

$$FP = \{F(x) | \, x \in CP\}$$

A Definição 1 formaliza o conceito de dominância entre as soluções, enquanto a Definição 2 apresenta os conceitos de Conjunto de Pareto e Fronteira de Pareto. Há diversos métodos que exploram o CP a fim de identificar a soluções pertencentes à FP, dentre os quais os algoritmos evolucionários multiobjetivo merecem destaque. Até novembro de 2015, aproximadamente 10.000 trabalhos foram publicados em computação evolucionária multiobjetivo[8].

Zhou et al. (2011) afirmam que, devido serem métodos baseados em população, algoritmos evolucionários são capazes de aproximar a FP de um POM em uma única execução, por isso o interesse crescente no uso destes algoritmos para esta classe de problema. Um algoritmo genético mono-objetivo pode ser modificado para achar o conjunto de múltiplas soluções não dominantes em uma única execução devido sua habilidade de explorar diferentes regiões do espaço de busca simultaneamente, além disso, o operador de cruzamento pode combinar boas

[8] Dados baseados no repositório http://delta.cs.cinvestav.mx/~ccoello/EMOO/EMOOstatistics.html, o qual é mantido pelo Professor Carlos A. Coello Coello. Acesso em: 10 de novembro de 2015.

soluções para diferentes objetivos, criando novas soluções não-dominadas em partes não exploradas da FP. Adicionalmente, a maior parte dos algoritmos genéticos multiobjetivos não requer que o usuário priorize ou atribua pesos às funções objetivos. Por estes motivos, os AG têm sido uma das heurísticas mais populares para lidar com problemas de otimização multiobjetivo (ZHOU et al., 2011).

Em adição às duas abordagens para lidar com POM acima mencionadas (a saber, transformação do problema multiobjetivo em um problema mono-objetivo e a baseada em Fronteira de Pareto), Freitas (2004) discute a abordagem baseada em lexicografia para lidar com otimização multiobjetivo na mineração de dados. A ideia básica desta abordagem é atribuir diferentes prioridades aos diferentes objetivos, otimizando-os conforme a ordem de prioridade.

No mesmo trabalho o autor discute os prós e os contras de cada uma das três abordagens supracitadas. Em resumo, apesar da abordagem convencional onde transforma-se o problema multiobjetivo em mono-objetivo por meio de ser a mais utilizada na literatura de mineração de dados, os argumentos que advogam contra a sua adoção em detrimento às abordagens baseadas em lexicografia e em fronteiras de Pareto incluem a dificuldade na atribuição de pesos e a existência de diferentes escalas/grandezas para as medidas a serem consideradas como função objetivo, o que pode produzir resultados enviesados.

A abordagem baseada em lexicografia pode ser considerada intermediária entre a transformação multiobjetivo em mono-objetivo e a baseada em fronteiras de Pareto. Nesta abordagem, o usuário apenas deve determinar o ranqueamento das funções objetivo e ela retorna uma única solução ao usuário, tal como a transformação multiobjetivo em mono-objetivo, mas sem a necessidade de atribuição de pesos ou de normalização de diferentes escalas/grandezas. Já a abordagem baseada em fronteiras de Pareto retornam um conjunto de soluções não dominantes, que para ser encontrado, requer um custo computacional maior que as outras duas abordagens, mas que oferecem ao usuário uma rica fonte de informação sobre a relação entre as funções objetivos. Adicionalmente, as diferentes soluções podem ser combinadas de forma a gerar uma solução mais robusta (*e.g.* usando conjunto de classificadores) (FREITAS, 2004).

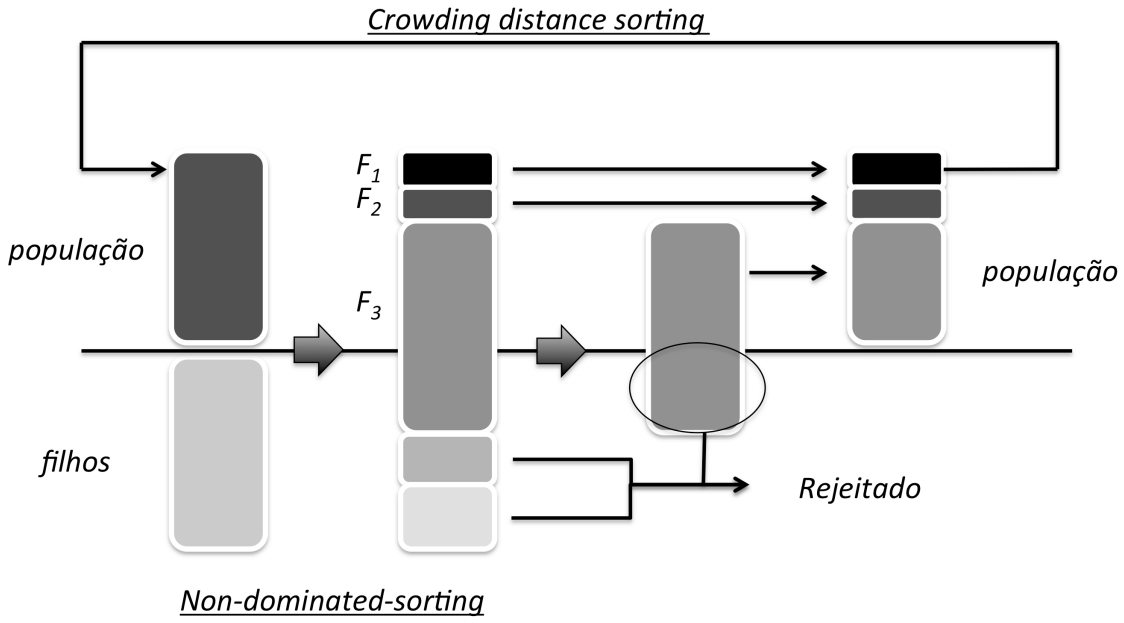
Diversos algoritmos evolucionários têm sido propostos para explorar as fronteiras de Pareto, dentre as tendências na área destacam-se a utilização de métodos híbridos e de paralelismo a fim de reduzir o custo computacional (Coello Coello, 2015), em ambas as tendências os algoritmos genéticos encontram-se presentes por serem facilmente paralelizáveis e adaptáveis (ZHI-XIN; JU, 2009). Konak, Coit e Smith (2006) apresentam um tutorial sobre otimização multiobjetivo usando algoritmos genéticos onde comparam 13 métodos bem estabelecidos, dentre os quais o *Vector Evaluated Genetic Algorithm* (VEGA) (SCHAFFER, 1985), *Nondominated Sorting Genetic Algorithm* (NSGA) (SRINIVAS; DEB, 1994), Fast Nondominated Sorting Genetic Algorithm (NSGA-II) (DEB et al., 2002), *Strength Pareto Evolutionary Algorithm* (SPEA) (ZITZLER; THIELE, 1999), *improved* SPEA (SPEA2) (ZITZLER; LAUMANNS; THIELE, 2001) *etc*; sendo que o NSGA-II é um dos mais utilizados. Fato também

observado por Mukhopadhyay et al. (2014a) e Mukhopadhyay et al. (2014b) no contexto da mineração de dados, por exemplo, em um comparativo de algoritmos evolucionários multiobjetivos aplicados à classificação, 10 de 14 trabalhos utilizaram o NSGA-II como algoritmo-base, o que perfaz aproximadamente 70% dos estudos analisados pelos autores supracitados.

Dentre as justificativas para adoção deste algoritmo, destacam-se a sua eficiência computacional, a característica elitista e a facilidade de paralelização. O NSGA-II implementa o conceito de dominância por meio da ordenação das soluções nas fronteiras de Pareto, o que é feito por dois algoritmos de ordenamento, o *non-dominated sorted algorithm* e o *crowding distance sorting* aplicados à população conjunta (pais e filhos). Estes algoritmos buscam soluções próximas à frente de Pareto (que contém os indivíduos mais aptos) e soluções distribuídas no espaço, respectivamente.

No *non-dominated sorted algorithm*, para cada indivíduo, verificam-se quantas soluções a dominam e quais são por ela dominadas, então retiram-se as soluções não dominadas e diminui-se o contador das soluções dominadas pelas que foram retiradas, portanto, a cada etapa uma nova frente de Pareto é criada. O segundo procedimento do NSGA-II calcula a distância entre soluções que pertencem a uma mesma fronteira visando garantir uma melhor aptidão aos indivíduos que estejam em regiões menos povoadas, conferindo diversidade às soluções. A Figura 8 resume graficamente o funcionamento destes operadores.

Figura 8 – Desenho esquemático do funcionamento dos operadores do NSGA-II.



2.4.2 PROGRAMAÇÃO GENÉTICA

A programação genética, semelhante aos algoritmos genéticos, é uma técnica de computação evolucionária que resolve problemas de forma automática sem requerer que o usuário conheça ou especifique a forma ou estrutura da solução antecipadamente (POLI; LANGDON; MCPHEE, 2008). Inicialmente idealizada para encontrar programas de computador (BANZHAF et al., 1998), atualmente é empregada também na extração de padrões de bases de dados, pois é uma heurística flexível que permite a representação de padrões complexos (ESPEJO; VENTURA; HERRERA, 2010).

Esta técnica segue o funcionamento básico de algoritmos evolucionários descrito na Figura 5, seu principal diferencial diz respeito à codificação do indivíduo que comumente é representado por **árvores sintáticas**, onde os **nós internos** podem assumir funções matemáticas a partir de um conjunto pré definido, e as **folhas** representam as variáveis terminais. A união dos conjuntos de funções e de terminais formam o conjunto chamado de **primitivas** de um sistema de programação genética (POLI; LANGDON; MCPHEE, 2008). O Quadro 2.3 mostra exemplos de primitivas nos conjuntos de funções e variáveis terminais.

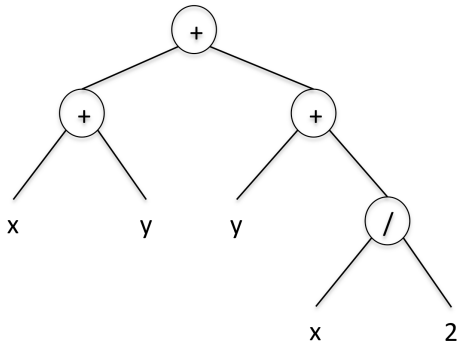
Quadro 2.3 – Exemplo de primitivas em funções e terminais de programação genética.

Conjunto de Funções	
Tipo de Primitiva	Exemplo(s)
Aritmética	+, *, /
Matemática	sen, cos, exp
Lógicas	E, OU, NÃO
Condicional	SE-ENTÃO-SE NÃO
Repetitivas	PARA-REPETIR
Conjunto de Terminais	
Tipo de Primitiva	Exemplo(s)
Variáveis	x, y
Valores Constantes	3; 0,45

Fonte: Adaptada de Poli, Langdon e McPhee (2008).

Conforme visto no Quadro 2.3, é possível definir diversos tipos de funções, como aritméticas, trigonométricas, lógicas *etc*. Por este motivo a programação genética tem ganhado destaque como método de regressão. Por exemplo, é possível relacionar um atributo (dependente) de um conjunto de dados com outros atributos (ditos independentes) por meio de funções matemáticas. Neste contexto, a Figura 9 mostra a relação entre duas variáveis, x e y (atributos independentes) para obter o valor de z (atributo dependente).

Figura 9 – Codificação de um indivíduo na programação genética utilizando árvores.



Fonte: Adaptada de Koza (1992).

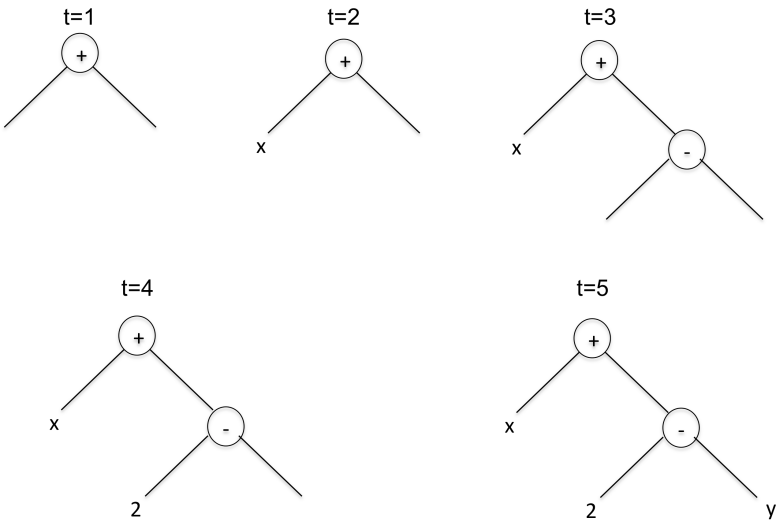
A Figura 9 exemplifica a codificação de um indivíduo como uma árvore binária - árvore que os nós possuem graus zero, um ou dois - resultando na expressão $(x+y)+(y+\frac{x}{2})$, a qual é a equação de regressão do atributo z. Tal como nos algoritmos genéticos, os operadores que implementam a inicialização, seleção, cruzamento e mutação, são dependentes do esquema de codificação adotado.

Por ser tratar de uma árvore, é necessário que sejam informados alguns parâmetros para inicializar os indivíduos (*e.g.* grau, altura máxima, se deve ser completa *etc*), o método de preenchimento. Um dos métodos de inicialização mais utilizados é o *ramped half-and-half*, que inicializa metade dos indivíduos pelo método de crescimento (*grow)* e a outra metade pelo método completo (*full*), isso auxilia a garantir que as árvores geradas tenham variabilidade de tamanhos e formas (KOZA, 1992).

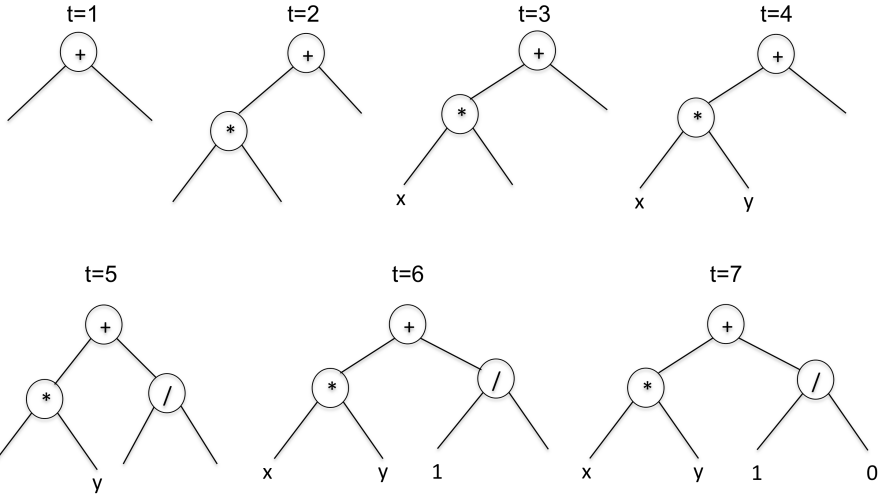
No método de inicialização completo, os nós são adicionados até gerarem uma árvore completa; já no método de crescimento os nós são adicionados até alcançarem um nó completo na altura máxima pré-definida, em ambos os métodos as funções e terminais são posicionados aleatoriamente, a altura da árvore de cada indivíduo também é definida aleatoriamente. A Figura 10 mostra o raciocínio dos métodos de inicialização.

A Figura 10a mostra a construção de uma árvore segundo o método *grow* com limite de altura igual a 2, o primeiro elemento gerado aleatoriamente é a raiz, cujo valor é "+", na próxima iteração, seleciona-se aleatoriamente o primeiro argumento da raiz, cujo valor escolhido foi "x'", por ser um terminal, este galho é fechado prevenindo que ele continue a crescer. O outro argumento escolhido (t=3) é uma função, por isso seus argumentos são forçados a serem terminais a fim de garantir que a árvore resultante não exceda a altura limite. O método de inicialização *full* é apresentado na Figura 10b, a escolha dos operadores e terminais segue o mesmo raciocínio do método *grow*, no entanto, este método garante que todas as folhas tenham a mesma altura.

Com os indivíduos inicializados, passa-se para as etapas de seleção, cruzamento e mutação. Os operadores de seleção são independentes do esquema de codificação do indivíduo, por



(a) Exemplo da inicialização de um indivíduo pelo método de crescimento.



(b) Exemplo da inicialização de um indivíduo pelo método completo.

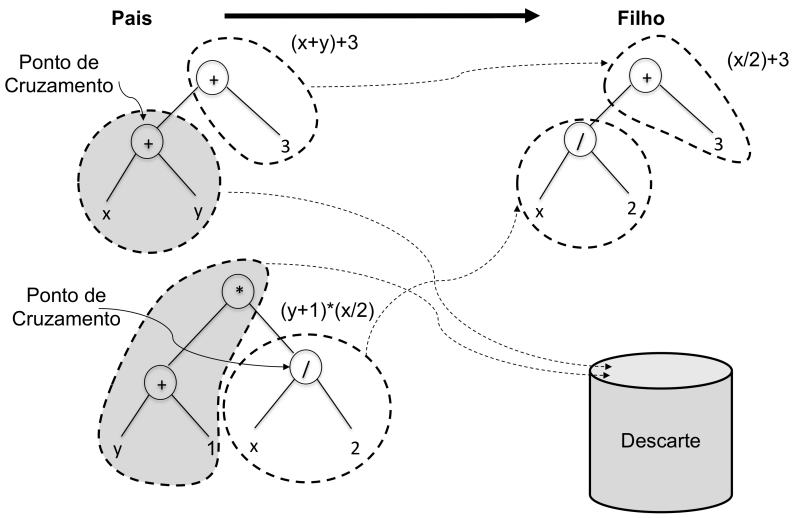
Figura 10 – Métodos *full* e *grow* de inicialização de indivíduos na programação genética.

Fonte: Adaptada de Poli, Langdon e McPhee (2008).

conseguinte, os operadores baseados em torneio ou roleta descritos para os algoritmos genéticos podem ser utilizados na programação genética. Já os operadores de cruzamento e mutação dependem do esquema de codificação e por isso diferem-se consideravelmente de outros algoritmos evolucionários.

Na programação genética, um dos operadores de cruzamento para indivíduos codificados em árvore mais utilizados é denominado de cruzamento de subárvore (*subtree crossover*). Dado dois parentes, este operador seleciona randomicamente um nó como o ponto de cruzamento em cada pai (tal como o *one-point crossover* do AG), duas subárvores (uma de cada pai) são então combinadas, gerando um único filho. As subárvores não selecionadas são descartadas. Um exemplo de cruzamento que ilustra este raciocínio é apresentado na Figura 11.

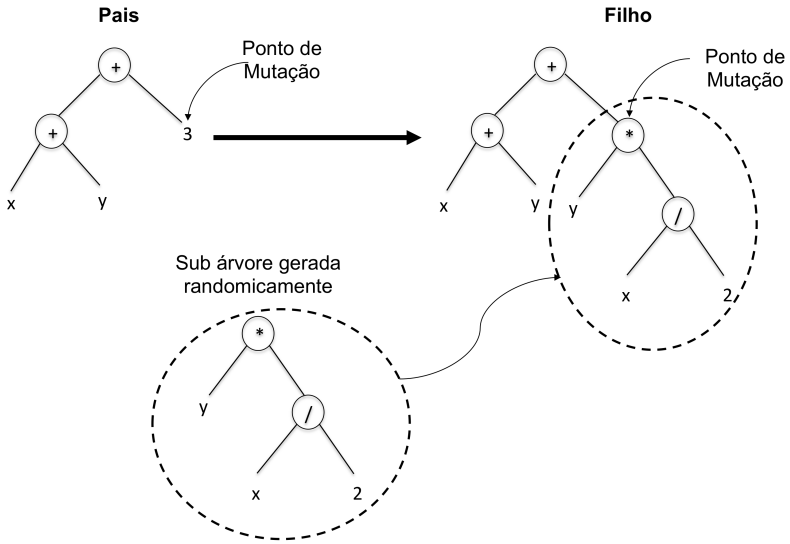
Figura 11 – Exemplo esquemático do operador de cruzamento *subtree crossover*.



Fonte: Adaptada de Poli, Langdon e McPhee (2008).

Tal como no cruzamento, o operador de mutação também é customizado para trabalhar com indivíduos codificados em árvore. Um dos mais utilizados é a mutação de subárvore (*subtree mutation*), que funciona de forma semelhante ao *subtree crossover*. Neste operador, um ponto de mutação é selecionado e a partir dele gera-se uma subárvore, ambas operações são efetuadas randomicamente. A Figura 12 exemplifica este raciocínio.

Figura 12 – Exemplo esquemático do operador de mutação *subtree mutation.*



Fonte: Adaptada de Poli, Langdon e McPhee (2008).

No exemplo dado pela Figura 12, o indivíduo selecionado para sofrer a mutação tem um

ponto selecionado aleatoriamente, no caso o filho à direita da raiz. Então, gera-se uma árvore aleatória que substituirá a subárvore selecionada no ponto de mutação, como resultado tem-se o indivíduo modificado.

2.5 CONSIDERAÇÕES FINAIS

Neste capítulo apresentou-se brevemente alguns temas relevantes para esta proposta de tese, iniciando pela contextualização da análise de dados, com foco no processo de extração de conhecimento de bases de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) – principalmente na descrição das principais etapas do processo (HAN; KAMBER, 2006). Posteriormente, descreveu-se um problema onipresente na análise de dados, a falta de dados em determinados exemplos (casos, instâncias), que possui diversas denominações na literatura como dados faltosos, dados faltantes, dados incompletos e valores ausentes (OLIVEIRA, 2009; SILVA, 2010; VERONEZE, 2011b; FACELI et al., 2011).

Neste âmbito, apresentou-se o panorama da área, destacando trabalhos relevantes como o de Little e Rubin (1987), marco nas pesquisas envolvendo VA – onde apresentam-se os três mecanismos de ausência de dados que, até hoje, atraem estudos relacionados. Outros conceitos pertinentes também foram apresentados, tais como o padrões de ausência de dados propostos por Schafer e Graham (2002) e quantificação de dados faltosos. Estas informações subsidiam o desenvolvimento de soluções para mitigar os efeitos nocivos desta problemática. De forma geral, é possível categorizar estes métodos em 4 classes; as abordagens ingênuas, também chamadas de métodos tradicionais ou de análise de caso completo; as de imputação de dados, que substituem o valor ausente por um valor estimado; as baseadas em modelos; e por último, os métodos baseados em aprendizado de máquina que visam o desenvolvimento/adaptação dos algoritmos com o intuito de aumentar sua robustez quanto a esta problemática (GARCÍA-LAENCINA; SANCHO-GóMEZ; FIGUEIRAS-VIDAL, 2009). Reitera-se que, qualquer que seja a abordagem utilizada, o objetivo principal é diminuir o viés imposto por esta problemática ao processo de análise de dados (GRAHAM, 2009).

Por fim, apresentou-se alguns conceitos de computação evolucionária, posicionando este campo de estudo no contexto da computação natural (De Castro, 2007), destacando o funcionamento básico de um algoritmo evolucionário e relacionando com seu paralelo na teoria da evolução, bem como suas características e aplicabilidade. Apresentou-se também dois algoritmos evolucionários, o algoritmo genético e a programação genética, discutindo-se particularidades quanto ao funcionamento, parametrização e, no caso dos algoritmos genéticos, sua aplicabilidade a problemas multiobjetivos.

3 DESCRIÇÃO DO PROBLEMA DO PROBLEMA

3.1 CONSIDERAÇÕES INICIAIS

Um problema pode ser avaliado por diversos prismas, sendo que a forma pela qual ele é observado influencia diretamente no desenvolvimento, reuso e melhorias de soluções. À vista disso, percebe-se que a descrição de um problema é necessária, suas bases residem na ciência formal, a qual preocupa-se com sistemas formais como lógica, matemática, estatística, teoria da computação, teoria da informação *etc*. Além de caracterizar o sistema em estudo, a ciência formal busca prover informações acerca das estruturas para uso posterior nas descrições dos fenômenos ou inferências (FRANKLIN, 1994).

Apesar dos avanços nas áreas em questão, há uma lacuna na literatura no tocante à descrição da imputação de dados como um problema de otimização combinatorial. Devido a este desafio em aberto, vislumbrou-se a possibilidade de se propor uma descrição do problema em questão. Como resultado, espera-se que a adoção da descrição proposta facilite a formalização do problema e consequente desenvolvimento e adaptação de estratégias de otimização aplicadas ao contexto do tratamento de valores ausentes, com isso, aumentando o número e a qualidade de métodos desenvolvidos subsequentemente.

3.2 PROBLEMAS DE OTIMIZAÇÃO COM VARIÁVEIS MISTAS

Diversos problemas de otimização do mundo real podem ser modelados utilizando combinação de variáveis contínuas e discretas, dentre os quais encaixam-se a mineração de dados e reconhecimento de padrão, consequentemente, a análise de dados com valores ausentes. Devido a relevância prática desta classe de problemas, diversas estratégias e métodos vêm sendo desenvolvidos para manipulá-los.

As variáveis discretas destes problemas podem ser ordinais ou categóricas, as ordinais apresentam relação de ordenamento (*e.g.* baixo, médio e alto) e normalmente são tratadas usando uma abordagem baseada em relaxamento contínuo (GUO et al., 2004; RAO; XIONG, 2004). Variáveis categóricas possuem valores contidos em um conjunto finito de categorias, as quais caracterizam atributos que não possuem ordem entre si, como raça e gênero, por exemplo. As variáveis numéricas representam grandezas mensuráveis e podem ser inteiras ou reais, como por exemplo, idade e altura, respectivamente. Liao et al. (2014) indicam que as abordagens para lidar com problemas com variáveis mistas disponíveis na literatura tratam misturas de variáveis contínuas e categóricas, ou seja, eles não consideram a formulação de um problema que envolva, ao mesmo tempo, os três tipos de variáveis. Devido à ampla diversidade de estratégias para estas combinações, a presente proposta de formulação da imputação de dados como um

problema de otimização trata da mistura contínua e categórica.

Como levantado no Capítulo 1, em conjuntos de dados com atributos mistos, o processo de imputação pode ser visto como um problema de otimização com variáveis mistas, um modelo para este tipo de problema pode ser definido como segue:

Definição 1: Um modelo $R = (\boldsymbol{S}, \Omega, f)$*, onde:*

- $\boldsymbol{S}$ *é o espaço de busca definido sobre variáveis de decisão discretas e contínuas;*
- Ω *é um conjunto de restrições em relação às variáveis de decisão; e*
- $f : \boldsymbol{S} \rightarrow \mathbb{R}_0^+$ *é a função objetivo a ser minimizada.*

Deste modo, a solução $S \in \mathbf{S}$ é um uma atribuição de valores considerados factíveis para as variáveis de decisão, pertencentes à S, que satisfaçam as restrições contidas no conjunto Ω. O ótimo global S^* é uma solução factível a qual minimiza a função objetivo e pertence ao conjunto de soluções globais $\mathbf{S}^*$. Resolver um problema deste tipo é encontrar pelo menos uma solução $S^* \in \mathbf{S}^*$.

3.3 IMPUTAÇÃO DE DADOS COMO UM PROBLEMA DE OTIMIZAÇÃO

Nesta seção o modelo formal proposto será apresentado. Algumas estratégias para representação das soluções candidatas e de busca, incluindo operadores de busca e inicialização, são discutidas; funções de avaliação dispostas na literatura de tratamento de valores ausentes também são apresentadas.

3.3.1 REPRESENTAÇÃO DAS SOLUÇÕES CANDIDATAS

Seja o conjunto de dados hipotético disposto no Quadro 3.1 que possui atributos discretos e contínuos, e valores ausentes. Os atributos At1 e At2 são contínuos e representam peso e altura, já os atributos At3 e At4 são discretos, sendo categóricos (gênero) e ordinais (classificação de massa corpórea), respectivamente.

Intuitivamente, atributos discretos são representados a partir do conjunto contendo todos os valores observados. Apesar da relação de ordem entre si existentes nos dados ordinais, o modelo proposto trata todos os atributos discretos como dados categóricos a fim de diminuir a complexidade. Já os atributos contínuos podem ser representados por funções, valores médios de determinadas faixas, como por exemplo histogramas ou funções de densidade aproximadas, ou ainda, podem também ser obtidos a partir do conjunto com todos os valores observados tal como nos atributos discretos. Dependendo da natureza dos dados (quantidade de amostras e de valores), da precisão requerida e da técnica de análise empregada, a representação dos dados é escolhida. Em relação à representação da solução candidata para imputação de dados mistos,

Quadro 3.1 – Conjunto de dados com dados discretos e contínuos.

ID	At1	At2	At3	At4
1	66	1,66	feminino	normal
2	90	1,80	feminino	acima
3	75	1,90	masculino	normal
4	45	1,60	feminino	abaixo
5	?	1,70	masculino	obeso
6	70	?	?	acima
7	91	1,82	?	acima
8	?	1,66	masculino	?
9	88	?	feminino	obeso
10	62	1,83	masculino	?

Fonte: Elaborada pelo autor.

geralmente os métodos de imputação combinam diferentes estratégias. Por exemplo, métodos de imputação baseados em *k*NN usam a o valor médio dos vizinhos para atributos numéricos e o valor mais frequente (moda) para atributos discretos (BATISTA; MONARD, 2003).

Analisando a predição de valores para substituir os dados ausentes como um problema de otimização combinatorial, cada dado ausente (ou conjunto de dados ausentes de um mesmo atributo) pode ser visto como uma omissão a ser otimizada, onde o objetivo é encontrar o valor mais próximo ao que foi omitido. Observam-se na literatura duas estratégias básicas para representar as soluções candidatas: i) tratar cada valor individualmente; e ii) agrupando as instâncias e imputando valores únicos para os atributos com VA para o conjunto de instâncias de um grupo.

Neste sentido, baseando-se o conjunto de dados apresentado no Quadro 3.1, tem-se cada atributo com dois valores ausentes, portanto, oito valores a serem encontrados. Se considerarmos os valores observados como possíveis valores a serem imputados, conforme proposto pela primeira estratégia, teremos 8, 7, 2 e 4 valores possíveis para os At1, At2, At3 e At4, respectivamente. Por conseguinte, o número total de combinações possíveis é igual a 200.704 (resultado de: 8 x 8 x 7 x 7 x 2 x 2 x 4 x 4). Esta estratégia é adotada por Figueroa García, Kalenatic e López Bello (2011) e Patil e Bichkar (2010), dentre outros.

A segunda estratégia consiste em agrupar instâncias semelhantes em termos de incidência de valores ausentes, diminuindo assim número de combinações possíveis. Por exemplo, é possível formar dois grupos com as instâncias com VA, o grupo 1 inclui instâncias 5, 8 e 10, pois possuem VA apenas nos atributos At1 e At4; e o grupo 2 é formado pelas instâncias 6, 7 e 9, já que possuem VA nos atributos At2 e At3. Sendo assim, o grupo 1 apresentará 32 combinações possíveis (8 x 4) e o grupo 2, 14 combinações (7 x 2), perfazendo um total de 448 combinações. Apesar de perder variabilidade e incluir um passo adicional para realizar o agrupamento de instâncias, tal estratégia reduz drasticamente o espaço de busca e mostra-se competitiva para

conjuntos de dados reais, conforme analisado em Lobato et al. (2015b). Esta estratégia também é adotada por Tran, Zhang e Andreae (2015), por de Andrade Silva e Hruschka (2009) e por abordagens *hot-deck* em Andridge e Little (2010), por exemplo.

3.3.2 ESTRATÉGIAS DE BUSCA E INICIALIZAÇÃO

A escolha da estratégia de busca e de inicialização é dependente, sobretudo, da representação das soluções candidatas e das medidas de desempenho adotadas, a maior parte delas utiliza apenas métodos de busca global, com destaque para os modelos bioinspirados, como algoritmos genéticos, colônia de formigas e otimização por enxame de partículas. Alguns trabalhos utilizam ainda redes neurais artificiais, todavia, devido ao padrão totalmente aleatório (MCAR) ser um dos mais estudados e pelas características da análise multivariada, onde a predição de valores para um determinado atributo baseando-se nos demais nem sempre é indicada dada

Em relação às estratégias de busca, os métodos de imputação dividem-se em duas categorias, as que investigam todo o espaço de busca e as que dividem os espaços de busca em subespaços e ali aplicam os algoritmos. A diferença entre elas reside basicamente no fato de que o segundo método agrupa instâncias e então aplica processos de busca nos grupos, podendo imputar o mesmo valor para um determinado atributo ou então tratá-las individualmente - dependendo da representação das soluções candidatas e da definição da espaço de busca. Diversos critérios são utilizados, diz-se que a estratégia de busca é independente da classe se não levar em consideração o(s) rótulo(s) para predizer os valores ausentes - tais métodos são preferíveis pois são adequados à imputação online, quando recebe-se uma instância para predição que apresenta valor ausente. Em contraste tem-se métodos dependentes dos rótulos, como o *Concept most common attribute value for symbolic attribute and concept average value for numeric attribute* (CMC) (GRZYMALA-BUSSE; HU, 2001), imputando a média ou moda nos atributos com valores ausentes, calculadas a partir das instâncias pertencentes a um mesmo rótulo - por analogia, é possível notar que métodos pertencentes a esta categoria não são apropriados para adoção em tempo de classificação.

Apesar de não observadas nos trabalhos envolvendo tratamento de valores ausentes usando modelos bioinspirados, é possível a utilização de restrições baseadas em conhecimento de fundo, como por exemplo é possível utilizar restrições dos tipos *Must-Link* e *Cannot-Link* (WAGSTAFF et al., 2001), a fim de evitar a imputação de valores espúrios, como associar em um mesmo exemplo valor “Homem” ao valor “Gravidez positiva” (BARALDI; ENDERS, 2010); ou ainda, adoção de restrições para a predição de valores negativos em atributos numéricos que possuem apenas valores positivos.

Em relação à inicialização, grande parte dos métodos de TVA que requerem este procedimento utilizam-se da inicialização aleatória, não somente os que tratam a imputação de dados como um problema de otimização. Por exemplo, métodos *hot deck* normalmente escolhem os

doadores de forma aleatória (ANDRIDGE; LITTLE, 2010); este fato é também observado em todos os algoritmos genéticos utilizados para a imputação de dados analisados no Capítulo 4. Alternativamente, alguns trabalhos, como Tran, Zhang e Andreae (2015) e Gautam e Ravi (2014), utilizam algum método de imputação simples para inicializar a solução. Esta estratégia também é investigada nesta tese, conforme será discutida adiante.

3.3.3 FUNÇÕES OBJETIVO

O projeto das funções objetivo, para maior parte dos casos, é dependente da tarefa de modelagem. Por exemplo, séries temporais possuem propriedades diferenciadas que inviabilizam o uso de processos convencionais de análise de dados (Figueroa García; KALENATIC; López Bello, 2010). São estas:

- Uma série temporal possui estruturas autocorrelacionadas;
- uma série temporal pode conter componentes sazonais ou de tendência;
- o objetivo principal em análise de séries temporais é predizê-las;
- a ergodicidade é um operador de defasagem importante.

À vista disso, estatísticas como média, variância e estruturas de autocorrelação são consideradas medidas úteis para avaliar o impacto da imputação de dados (HUNG, 2008; FLORES; COTA; MORALES, 2011). Variantes destas estatísticas, como a matriz de covariância e determinantes também são usados por alguns autores que utilizam técnicas de otimização para realizar a imputação de dados (DORRI; AZMI; DORRI, 2012; KRISHNA; RAVI, 2013; GAUTAM; RAVI, 2015). Usualmente, calcula-se a estatística para as instâncias completas e depois para o conjunto de dados imputado, sendo computada a diferença entre elas - neste caso, o objetivo é diminuir a diferença (erro associado). No que tange à classificação de padrões, a literatura aponta informações da construção do modelo como medidas preferíveis para avaliação do impacto da imputação (FARHANGFAR; KURGAN; DY, 2008; HRUSCHKA et al., 2009). Nesse horizonte, as seguintes medidas destacam-se (BARROS; BASGALUPP; CARVALHO, 2015):

- **Acurácia do modelo:** na imputação de dados no contexto de classificação de padrões é a medida mais observada, representa a taxa de instâncias corretamente classificadas e é calculada conforme a Eq. 3.1.

$$acuracia = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.1}$$

- **Precisão (*precision*) e sensitividade (*recall*):** a precisão denota a taxa de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas, já

a sensitividade é a taxa de amostras positivas classificadas corretamente sobre o total de amostras positivas. A sensitividade é útil para avaliar o modelo de classificação em conjuntos de dados desbalanceados. Ambas também compõem o cálculo de outras medidas de desempenho, a precisão e a sensitividade são calculadas de acordo com as Eq. 3.2 e 3.3, respectivamente.

$$precisao = \frac{tp}{tp + fp} \tag{3.2}$$

$$sensitividade = \frac{tp}{tp + fn} \tag{3.3}$$

- ***F-measure:*** mede a eficiência do modelo de classificação levando em consideração o erro, mesmo não levando em consideração os falso positivos (POWERS, 2011). O *F-measure* é calculado como a média harmônica da precisão e sensitividade (*recall*), conforme disposto na Eq. 3.4.

$$F1 = 2 \times \frac{precisao \times sensitividade}{precisao + sensitividade} \tag{3.4}$$

Outras medidas também podem ser utilizadas, como curvas ROC para classificadores de uma forma geral, ou ainda, a classificação exata e *Hamming Loss* no caso da classificação multirrótulo, por exemplo.

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo, apresentou-se um esboço da imputação de dados como um problema de otimização com o objetivo de estender as potencialidades da aplicação de algoritmos de busca bioinspiradas neste nicho de aplicação. Também discutiu-se elementos pertinentes à descriçãp/formalização como a representação das soluções, possíveis estratégias de busca, inicialização e funções objetivos adotadas.

Destaca-se que, apesar das estratégias serem discutidas de maneira isolada, as mesmas podem ser combinadas de acordo com as especificidades do domínio de aplicação. Por exemplo, é possível utilizar de métodos de imputação simples para inicialização das soluções candidatas e estratégias de busca baseadas em agrupamento concomitantemente. Já outras estratégias discutidas, como a incorporação de conhecimento de fundo por meio de restrições dos tipos *Must-Link* e *Cannot-Link* requer *expertise* da área relacionada ao conjunto de dados a fim de que tais restrições sejam extraídas, utilizadas e validadas. Os capítulos subsequentes analisam alguns cenários abordados aqui, como a utilização de subconjuntos para representar as soluções candidatas (Capítulo 5); a utilização de múltiplas funções objetivo (Capítulo 6); e a inicialização de soluções candidatas usando resultados de algoritmos de imputação simples (Capítulo 7).

4 TRABALHOS CORRELATOS

4.1 CONSIDERAÇÕES INICIAIS

O processo de aquisição e integração de dados é propício a falhas, sendo que um problema notório diz respeito aos valores ausentes, com efeitos nocivos para a análise dos dados. Por ser um problema ubíquo, capilar e de grande impacto, a literatura sobre valores ausentes é vasta, incluindo estudos acerca de metodologias de pesquisa e aquisição de dados de forma a evitar sua ocorrência (MCKNIGHT et al., 2007; NEWMAN, 2014); e estudos de como lidar com esta problemática em diversos domínios de aplicação como por exemplo em: estudos longitudinais (ENDERS, 2011; XIE, 2012; FERRO, 2014; YOUNG; JOHNSON, 2015), análise de séries temporais com valores ausentes (HUNG, 2008; CISMONDI et al., 2011; HONAKER; KING; KING, 2013; JUNGER; Ponce de Leon, 2015), processamento de sinais (HRYDZIUSZKO; VIANT, 2011) e classificação de padrões (BATISTA; MONARD, 2003; FARHANGFAR; KURGAN; DY, 2008; HRUSCHKA et al., 2009; NANNI; LUMINI; BRAHNAM, 2012; TRAN; ANDREAE; ZHANG, 2015).

Devido à grande abrangência da área, nesta seção serão analisados os trabalhos pertencentes ao escopo da tese: a classificação de padrões com valores ausentes e a imputação de dados, com foco nos métodos evolucionários. Dessa forma, o presente capítulo apresenta um panorama dos estudos relacionados ao tema, explicitando os trabalhos julgados mais relevantes, primeiramente analisando os desafios e tendências na área; para então analisar os trabalhos que utilizam-se da computação evolucionária no tratamento de valores ausentes.

4.2 REVISÕES DA LITERATURA

Como dito, o estudo dos dados ausentes remete à década de 70, por conseguinte, o número de trabalhos na área é grande. Por este motivo, a pesquisa inicial deu-se por revisões que apresentassem uma síntese dos métodos até então desenvolvidos. Em Schafer e Graham (2002) os autores realizam uma revisão dos métodos do estado da arte, levantando alguns pontos que até então permaneciam sem solução; direcionando assim, as pesquisas seguintes. Este trabalho foi atualizado por Graham (2009), contemplando um embasamento teórico maior e contextualizando com problemas reais, nele o autor também apontou um problema que permeia as publicações recentes: a não utilização de métodos de tratamento recentemente propostos.

Ainda em 2009, García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009) conduziram uma revisão sobre a classificação de padrões com valores ausentes, onde os autores apresentaram e compararam métodos de TVA já consolidados na área, mais especificamente: *k-Nearest Neighbor* (KNN); *Multi-Layer Perceptron* (MLP), *Self-organizing map* e *expectation–maximization*

algorithm. Por padrão, os autores definem os três tipos de mecanismos de dados faltosos, o diferencial reside na apresentação de dois cenários possíveis quando se trabalha com aprendizado de máquina neste contexto: a) o conjunto de dados utilizado para o treino é completo, enquanto os VA estão apenas no conjunto de teste; ou b) há ocorrência de VA em ambos os conjuntos de dados.

No que tange à avaliação do desempenho, há diferentes abordagens para imputação em bases para tarefa de classificação. García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009) consideram dois tipos de tarefas a serem realizadas: classificação e imputação. No primeiro, uma vez que os VA foram imputados, um classificador é treinado e sua acurácia é medida por meio da taxa de erro de classificação. Quanto à comparação dos métodos de imputação, dois critérios foram utilizados:

- Acurácia Preditiva (PAC): um método de imputação deve preservar os valores reais o máximo possível. Considerando que o i-nésimo atributo tem valores ausentes em alguns padrões de entrada, sua versão imputada $\tilde{X}_i$ deve ser próxima a $\hat{X}_i$ - dada a necessidade de saber o valor real $\hat{X}_i$, infere-se que esta medida pode ser utilizada apenas em bases sintéticas. A correlação de *Pearson* entre $\tilde{X}_i$ e $\hat{X}_i$ fornece uma boa medição da performance de imputação, e é dada por:

$$PAC \equiv r = \frac{\sum_{n=1}^{N}(\widetilde{X}_{i,n} - \bar{\widetilde{X}}_i)(\widehat{X}_{i,n} - \bar{\widehat{X}}_i)}{\sqrt{\sum_{n=1}^{N}(\widetilde{X}_{i,n} - \bar{\widetilde{X}}_i)^2 - (\widehat{X}_{i,n} - \bar{\widehat{X}}_i)^2}} \tag{4.1}$$

  onde $\tilde{X}_{i,n}$ e $\hat{X}_{i,n}$ denotam, respectivamente, o enésimo valor de $\tilde{X}_i$ e $\hat{X}_i$; e ainda, $\overline{\tilde{X}_i}$ e $\overline{\hat{X}_i}$ a média dos N valores inclusos em $\tilde{X}_i$ e $\hat{X}_i$. Um bom método de imputação gerará para a correlação de *Pearson* um valor próximo de 1, denotando um pequeno afastamento dos valores reais.

- Acurácia Distributiva (DAC): um método de imputação deve preservar a distribuição dos valores reais. Uma medida de preservação da distribuição destes valores é a distância entre a função de distribuição empírica, tanto para os valores imputados quanto para os reais. As funções de distribuição empíricas, $F_{\hat{X}_i}$ para os casos com valores reais, e $F_{\tilde{X}_i}$ para casos com valores imputados, são definidas por:

$$F_{\widehat{X}_i}(X) = \frac{1}{N}\sum_{n=1}^{N} I(\widehat{X}_{i,n} \leq X) \tag{4.2}$$

$$F_{\widetilde{X}_i}(X) = \frac{1}{N}\sum_{n=1}^{N} I(\widetilde{X}_{i,n} \leq X) \tag{4.3}$$

o I é a função indicadora. A distância entre essas funções pode ser medida por meio da aplicação do teste de *Kolmogorov-Smirnov*, D_{KS}, que é dada por:

$$DAC \equiv D_{KS} = \max_{n} (\|F_{\widehat{X}_i}(X_n) - F_{\widetilde{X}_i}(X_n)\|) \tag{4.4}$$

os X_n valores são o conjunto ordenado de valores verdadeiros e imputados do atributo x_i. Quanto menor o valor desta distância, menor é o viés imposto, logo, melhor o método de imputação.

A escolha dos métodos de imputação feita por García-Laencina, Sancho-Gómez e Figueiras-Vidal (2009) foi devido aos softwares comerciais de suporte à decisão não lidarem com dados incompletos, implicando na necessidade de tratá-los externamente. Os resultados experimentais permitiram concluir que não existe uma única solução que oferece resultados ótimos em cada domínio de aplicação. Por fim, os autores afirmam que, geralmente, em cenários reais é necessário realizar um estudo detalhado a fim de avaliar qual método de estimação pode propiciar um ganho de desempenho na classificação.

Wohlrab e Fürnkranz (2010) também conduziram uma revisão no contexto de classificação, mas específica ao paradigma de aprendizado de regras de associação, por conseguinte, algumas estratégias analisadas tais como a "*Any value strategy*", que trata os valores ausentes como "*don't care*"; a "*Pessimistic value strategy*", a qual pode ser vista como uma adaptação da "*reduced information gain*" (abordagem utilizada em árvores de decisão proposta por Quinlan (1989)); podem apenas ser utilizadas no aprendizado de regras de associação. Em resumo, os autores concluem que as abordagens baseadas em imputação de dados (as quais os autores denominaram *"Distributed" e "Predicted"*), mantêm a potencial preferência para atributos com valores ausentes, e que a combinação apropriada de estratégias com propriedades complementares pode ser uma abordagem promissora para alcançar bons resultados de forma confiável. Neste trabalho, Wohlrab e Fürnkranz (2010) também levantam o questionamento acerca das estratégias que são dependentes das classes, que consequentemente não podem ser adotadas em tempo de classificação, conforme será discutido adiante.

Ainda na discussão das revisões, Eekhout et al. (2012) conduziram uma revisão sistemática acerca de como os dados ausentes são reportados e manipulados na epidemiologia. Apesar de ser específica a uma área e ao ano de 2010, um dado interessante é que dos 262 trabalhos analisados, 81% reportaram análise de caso completo e 14% utilizaram-se de imputação simples. Portanto, mesmo com os avanços nos métodos de TVA propostos na literatura de estatística e de aprendizado de máquina, observa-se a sua não adoção pelos analistas. Fato que também vem sendo observado na psicologia (SCHAFER; GRAHAM, 2002; GRAHAM, 2009).

Cheema (2014) realizou uma revisão acerca dos métodos para manipular VA em pesquisas na área de educação, semelhante à de Eekhout et al. (2012). Nesta revisão, Cheema (2014)

contextualizou a problemática de ausência de dados nas pesquisas em educação além de comparar métodos para lidar com VA bem estabelecidos (*e.g. listwise/parwise deletion*, imputação por média, por regressão, *hot-deck*, zero, *single random*, *Last Observed Value Carried Forward Maximum Likelihood Expectation-Maximization* e imputação múltipla. Apesar da realização do estudo comparativo, as análises foram inconclusivas. Contudo, alguns direcionamentos podem ser retirados dessa revisão, como: i) a dificuldade de se comparar os métodos recentemente propostos, pois eles utilizam-se de diferentes metodologias experimentais, o que impossibilita a construção de *guidelines* que possam auxiliar na escolha da melhor forma de tratar o problema em questão; ii) necessidade de parametrização dos métodos, o que desencoraja sua adoção, e ainda, por falta de conhecimento específico, os analistas que optam por sua utilização, usam-no com os valores padrões para os parâmetros, o que pode inserir um viés desnecessário nas análises subsequentes.

Convém pontuar que a melhor forma de lidar com os VA depende de um conjunto de fatores como a natureza do conjunto de dados (tipos dos atributos, dimensionalidade, informações acerca da ausência dos dados *etc*) e os algoritmos a serem utilizados na fase de análise. Geralmente, os pesquisadores de determinadas áreas como educação e saúde não possuem *expertise* necessária para identificar e implementar o melhor método aplicável aos requisitos do problema (MCKNIGHT et al., 2007). Por este motivo, pesquisadores que não são familiarizados com métodos quantitativos talvez não estejam aptos a escolher o melhor método para lidar com valores ausentes nas suas próprias pesquisas (ENDERS, 2010).

4.2.1 CONSIDERAÇÕES DE ANÁLISE

Esta subseção resume alguns dos pontos resultantes das análises das revisões da literatura acima discutidas e que foram julgados pertinentes, são eles:

- A imputação de dados não é, exclusivamente, a melhor forma de lidar com dados faltosos. Na maior parte dos problemas reais é necessário a condução de um estudo detalhado sobre as características da ausência dos dados e suas proporções, a fim de que se escolha a melhor combinação dos métodos disponíveis e aplicáveis ao problema;
- Observa-se a não adoção de métodos de tratamento de valores ausentes na literatura de análise de dados, sobretudo em em pesquisas na área de educação e saúde. As hipóteses para este problema são:
  - A adoção de diferentes metodologias experimentais dificulta a comparação dos métodos propostos e impossibilita a construção de *guidelines* para auxiliar na escolha dos métodos de TVA;
  - A necessidade de parametrização da maior parte dos métodos desencoraja pesquisadores sem experiência na análise de dados com valores ausentes;

- No contexto aprendizado supervisionado, é necessário classificar os métodos de imputação em relação à dependência aos rótulos, pois métodos ditos rótulo-dependentes não são passíveis de utilização em tempo de classificação/regressão.

4.3 ESTUDOS COMPARATIVOS

Posicionadas as revisões sobre as formas de lidar com valores ausentes em diversas áreas, passa-se para a análise de trabalhos que realizaram comparações de métodos de imputação, onde percebe-se que o estudo dos diversos métodos de tratamento de valores ausentes em classificação é uma tarefa recorrente. Um dos primeiros trabalhos pertencentes a este escopo é o de Mundfrom et al. (1998), onde os autores observaram o impacto da imputação de dados em modelos de classificação. Para tal os experimentos foram realizados da seguinte forma: os modelos de classificação foram construídos com base nos dados completos, introduziram-se valores ausentes e aplicou-se a imputação por média, por regressão e usando *hot-deck*, então utilizou-se as bases imputadas para se medir o efeito na acurácia dos classificadores. Neste trabalho a imputação por média e a usando *hot-deck* saiu-se melhor do que a imputação por regressão. Em Grzymala-Busse e Hu (2001) os autores analisaram diversos métodos de imputação aplicados ao aprendizado de máquina, dentre os métodos de imputação analisados, destacam-se o *Concept Most Common Attribute Value*, o *Event-Covering* e o *Ignore Missing*, métodos até então adotados. Batista e Monard (2003) compararam quatro abordagens diferentes, duas das quais estão incorporadas internamente nos algoritmos C4.5 e CN2, e as outras duas são a imputação por média/moda e a imputação por k-Nearest Neighbor; Acuna e Rodriguez (2004) desenvolveram um trabalho na mesma direção, avaliando o impacto de quatro métodos de imputação, (deleção, imputação por média, imputação por mediana e KNNI).

Tanto Batista e Monard (2003) quanto Acuna e Rodriguez (2004) são de grande relevância para área pois os experimentos mostraram a eficiência do KNNI, guiando as pesquisas subsequentes. É importante notar que, à época dos trabalhos em questão, os métodos de tratamento de valores ausentes baseavam-se na análise estatística, sendo o KNNI um dos primeiros métodos baseados em aprendizado de máquina. Acuna e Rodriguez (2004) apontam as seguintes vantagens da imputação de dados por KNN: i) ela pode prever tanto atributos quantitativos quanto qualitativos; ii) ela não requer a criação de um modelo de predição para cada atributo com dado ausente; iii) é possível tratar instâncias com múltiplos valores ausentes; iv) este método leva em consideração a correlação entre os atributos. Como desvantagens, o autor cita: i) a escolha da medida de distância; ii) o mecanismo de busca do algoritmo requer que todo o conjunto de dados seja percorrido de forma a encontrar as instâncias mais semelhantes, o que pode ser crítico na mineração de dados em grandes bases; iii) o desempenho é sensível ao número de vizinhos definidos, tanto em termos de precisão quanto em custo computacional.

Farhangfar, Kurgan e Dy (2008) avaliaram o impacto de cinco métodos de imputação na classificação, o diferencial deste trabalho é que os autores restringiram ao domínio discreto.

Como conclusão, os autores afirmam que não há um método de imputação universalmente melhor para todos os classificadores. O diferencial deste estudo foi que os autores correlacionaram um método de imputação que melhor se aplica a um classificador como por exemplo, a combinação da imputação baseada em Naïve-Bayes com o classificador RIPPER para conjuntos de dados com grande quantidade de valores ausentes (entre 40% e 50%).

No mesmo sentido, o trabalho de Luengo, García e Herrera (2012) se faz notório, pois os autores discutem a correlação de um melhor método de imputação para um classificador/grupo de classificadores. Para tal, os autores confrontam 14 diferentes métodos de imputação, utilizando 23 métodos de classificação, os quais dividiam-se em três categorias: aprendizagem por indução de regras; métodos baseados em otimização e métodos baseados em distância. O principal parâmetro de avaliação utilizado foi a acurácia do modelo preditivo. Como principal contribuição deste estudo tem-se a correlação de qual método de imputação é mais aplicável para um determinado grupo de classificadores. Isto é realizado por meio do teste de hipótese *Wilcoxon signed-rank* (WILCOXON, 1945). Outros pontos relevantes remetem à análise da influência do método de imputação nos dados em relação a duas medidas: *Wilson's Noise Ratio* (WNR) e *average mutual information difference*, as quais podem ser definidas como:

- *Wilson's noise ratio*: esta medida proposta por Wilson (1972) observa o ruído no conjunto de dados. Para cada instância de interesse, o método procura os K vizinhos mais próximos, por meio da distância euclidiana, e utiliza os rótulos de classe de tais vizinhos. Isto é feito a fim de classificar o exemplo considerado: caso o exemplo não seja corretamente classificado, então o ruído é aumentado em uma unidade. Portanto, a relação de ruído final pode ser computada por meio da Eq. 4.5. Em geral, utiliza-se valores ímpares para k, (*e.g.* 3, 5 ou 7) vizinhos, de forma a evitar empates (FACELI et al., 2011).

$$Wilson's\ Noise\ Ratio = \frac{ruido}{n} \tag{4.5}$$

- *Average mutual information difference*: também chamada de *a mutual information*, é utilizada para indicar a relevância entre duas variáveis aleatórias (COVER; THOMAS, 1991), tornando-se conhecida no âmbito de seleção de atributos – correlacionando-os com a classe (KWAK; CHOI, 2002). Luengo, García e Herrera (2012) calcula esta medida de avaliação entre o atributo de entrada e a classe, obtendo um conjunto de valores, um para cada atributo de entrada. No passo seguinte, calcula-se a razão entre cada um destes valores, considerando o conjunto de dados imputado em relação ao conjunto de dados não imputado, portanto, a taxa destas razões mostra se a imputação de dados resultou em um ganho de informação, conforme segue:

$$Avg.\ MI\ Ratio = \frac{\sum_{xi \in X} \frac{MI_{\alpha}(x_i)+1}{MI(x_i)+1}}{|X|} \tag{4.6}$$

onde X é o conjunto de atributos de entrada, $MI_{\alpha}(i)$ representa o valor da MI do i-ésimo atributo do conjunto de dados imputado, e $MI(i)$ é o valor da MI do i-ésimo atributo de entrada do conjunto de dados não imputado. A correção de Laplace também foi aplicada, somando 1 tanto no denominador quanto numerador, uma vez que o valor de MI pode ser igual a zero para alguns atributos de entrada.

O Cálculo do $MI(x_i)$ depende do tipo do atributo x_i. Se o atributo é nominal, o MI entre X_i e o rótulo da classe Y é computado como segue:

$$MI_{nominal}(x_i) = I(x_i;Y) = \sum_{z \in x_i} \sum_{y \in Y} p(z,y)\, log_2 \frac{p(z,y)}{p(z)p(y)} \tag{4.7}$$

Por outro lado, se o atributo x_i é numérico, então usa-se a Janela de Parzen para estimar densidade, considerando a função da Janela Gaussiana.

$$MI_{numeric}(x_i) = I(x_i;Y) = H(Y) - H(C|X) \tag{4.8}$$

sendo que $H(Y)$ é a entropia para o rótulo da classe:

$$H(Y) = -\sum_{y \in Y} p(y)\, log_2\, p(y) \tag{4.9}$$

e $H(C|X)$ é a entropia condicional

$$H(Y|x_i) = -\sum_{z \in x_i} \sum_{y \in Y} p(z,y)\, log_2\, p(y|z) \tag{4.10}$$

Considerando que cada amostra possui a mesma probabilidade, aplicando-se a regra de Bayes e aproximando $p(y|z)$ por meio da Janela de Parzen, obtem-se:

$$\hat{H}(Y|x_i) = -\sum_{j=1}^{n} \frac{1}{n} \sum_{y=1}^{n} \hat{p}(y|z_j)\, log_2\, \hat{p}(y|z_j) \tag{4.11}$$

onde n é o número de instâncias do conjunto de dados, N é o número total de rótulos das classes e $\hat{p}(c|x)$ é

$$\hat{p}(y|z) = \frac{\sum_{i \in I_c} \exp(-\frac{(z-z_i)\sum^{-1}(z-z_i)}{2h^2})}{\sum_{k=1}^{N} \sum_{i \in I_k} \exp(-\frac{(z-z_i)\sum^{-1}(z-z_i)}{2h^2})} \tag{4.12}$$

Neste caso, I_c é o índice dos exemplos de treinamento pertencentes à classe c, e $\sum$ é a covariância da variável aleatória $(z - z_i)$.

Avaliar os métodos de imputação usando o Wilson's noise ratio permite observar qual método de imputação reduz o impacto dos valores ausentes como um ruído e quais métodos produzem ruído quando imputam. Já o uso do MI ocorre em dois sentidos. Alguns trabalhos a utilizam para avaliar a influência do método de imputação nas instâncias e atributos, individualmente, tal como em Luengo, Sáez e Herrera (2011), Luengo, García e Herrera (2012); ou então auxiliando no TVA, tal como Doquire e Verleysen (2012) que faz a seleção de atributos com valores ausentes usando estimadores de *mutual information*; ou outros estudos que utilizam esta medida para guiar o processo de imputação (KWAK; CHOI, 2002; GARCÍA-LAENCINA et al., 2009; PAN et al., 2015). Além das medidas de desempenho supramencionadas, o trabalho de Luengo, García e Herrera (2012) forneceu as bases experimentais para este projeto de tese, incluindo conjuntos de dados, métodos de imputação, classificadores, medidas de desempenho e metodologia de avaliação.

Outros estudos utilizam outras medidas de desempenho tal como a matriz de covariância, por exemplo, Liu e Brown (2013) comparam cinco métodos de imputação iterativos no contexto de classificação multivariada utilizando dois critérios de desempenho, um baseado na matriz de covariância e outro na acurácia do classificador antes e depois da imputação de dados. Novamente, nenhum método de imputação aparece como o melhor em todos os casos examinados. E ainda, os autores destacam que em um conjunto de dados com elevada proporção de valores ausentes, os danos às propriedades estatísticas são irreversíveis, o que dificulta a avaliação do processo de imputação. Como principal contribuição deste trabalho, ressalta-se que a otimização do critério baseado na covariância não necessariamente implica em uma melhora no critério baseado na acurácia do classificador.

Silva e Hruschka (2013) conduziram um estudo experimental acerca do uso de algoritmos de imputação baseados em vizinhos mais próximos aplicados à classificação, cujas perguntas de pesquisa foram: i) O desempenho dos algoritmos estudados são significativamente diferentes para uma combinação de mecanismos de ausência e taxas de ausência? ii) Qual a correlação entre medidas baseadas na acurácia preditiva da imputação (distância entre o valor real e o valor imputado calculado pelo erro quadrático médio (Root Mean Square Error (RMSE)) e a acurácia do classificador.

No tocante ao primeiro questionamento, o método *Iterative KNNI*, proposto por Brás e Menezes (2007), foi o que obteve melhores resultados, tanto para a acurácia preditiva da imputação quanto à acurácia do classificador; tal conclusão corrobora a potencialidade dos métodos de imputação iterativos, os quais refinam os estimadores parciais iterativamente até alcançar um estimador mais robusto - este projeto de tese baseia-se neste paradigma de imputação. Em relação à segunda questão de pesquisa, os autores observaram uma baixa correlação entre a acurácia dos métodos de imputação com a acurácia do classificador, portanto, os resultados sugerem que valores de RMSE considerados bons não levam necessariamente a bons resultados em termos da acurácia do classificador.

Este fato também foi observado em estudos dedicados a avaliar o impacto da imputação de dados na classificação e clusterização envolvendo análise de expressão gênica (AITTOKALLIO, 2010; LIEW; LAW; YAN, 2011). Segundo Souto, Jaskowiak e Costa (2015), é evidente que a avaliação do sucesso do tratamento de valores ausentes seja feita de forma prática, com foco no sucesso da estimação dos valores de expressão. Pode-se considerar, por exemplo, a habilidade do método de preservar genes significantes no conjunto de dados, ou seu poder preditivo/discriminativo para propósitos de agrupamento/classificação.

Tran, Andreae e Zhang (2015) avaliaram o impacto da imputação de dados na construção múltipla de atributos baseada em programação genética, também no contexto de classificação. Os objetivos da pesquisa eram: i) verificar o impacto da construção múltipla de atributos na acurácia de classificação usando dados imputados; ii) avaliar o efeito de cinco métodos de imputação na classificação usando atributos construídos. Como conclusões, os autores verificaram que a construção múltipla de atributos com programação genética melhora a acurácia de classificação para conjuntos de dados imputados e que possuíam poucos valores ausentes. Em relação à segunda pergunta de pesquisa, a combinação da construção de atributos com o método de imputação baseado em equações encadeadas (BUUREN; GROOTHUIS-OUDSHOORN, 2011) obteve os melhores resultados em relação à acurácia do classificador - uma ressalva deve ser feita em relação aos cinco conjuntos de dados utilizados nos experimentos pois eles possuíam apenas atributos numéricos.

Por meio da análise dos trabalhos discutidos acima, é possível perceber uma tendência no estudo da correlação entre as características da ausência de dados, métodos de imputação e algoritmos de classificação. Nesta direção, Sim, Kwon e Lee (2015) apresentam um método automático, que seleciona de forma adaptativa a combinação ótima do par algoritmo de classificação e método de imputação, considerando características do conjunto de dados. Apesar da semelhança com os trabalhos mencionados anteriormente, o diferencial deste estudo é a capacidade do método de identificar mudanças nas características do conjunto de dados em tempo real, e mudar a combinação do par imputação-classificação em tempo de execução. Isto é feito por meio de um mecanismo de raciocínio baseado em casos. Devido a sensibilidade à combinação entre conjunto de dados, algoritmo de classificação e método de imputação, os métodos propostos na tese foram desenvolvidos visando considerar diversas categorias de classificadores (*e.g. aprendizado de regras de associação, aprendizado baseado em instâncias etc*) em uma única solução, visando aumentar sua aplicabilidade em cenários reais.

Outros trabalhos que conduziram estudos comparativos têm em vista contribuições em um domínio de aplicação específico. Por exemplo, Chang e Ge (2011) comparam técnicas de imputação aplicadas a fluxo de dados; Ding e Ross (2012) comparam métodos para tratamento de dados faltosos em sistemas multibiométricos por meio de imputação; Peyre, Leplège e Coste (2011) realizam um comparativo, via simulação, de métodos de TVA aplicados a pesquisas sobre qualidade de vida; Yozgatligil et al. (2013) conduzem um estudo comparativo aplicado à

análise de série temporal-espacial, usando como estudo de caso dados meteorológicos; Gómez-Carracedo et al. (2014) comparam métodos de imputação simples aplicados à análise da qualidade do ar; e Inman, Elmore e Bush (2015) analisam dois métodos de imputação estatísticos no cenário de construção de demanda elétrica.

4.3.1 CONSIDERAÇÕES DE ANÁLISE

- Taxas de valores ausentes entre 1 e 5% são consideradas baixas e provocam pouco impacto na acurácia do classificador (ACUNA; RODRIGUEZ, 2004). No entanto, em alguns nichos de aplicação isso se torna crítico, como na análise de expressão gênica, onde 5% de valores ausentes podem afetar até 90% dos genes (SOUTO; JASKOWIAK; COSTA, 2015). Neste contexto, a imputação de dados pode ser uma ferramenta viável para diminuir o impacto desta problemática.

- Apesar de utilizadas inicialmente, provou-se que medidas de avaliação dos métodos de imputação baseadas em covariância ou no erro entre o dado imputado e o dado real não necessariamente estão relacionadas com um aumento na acurácia do classificador (LIU; BROWN, 2013; SILVA; HRUSCHKA, 2013; SOUTO; JASKOWIAK; COSTA, 2015);

- Um consenso nos trabalhos que conduziram estudos comparativos é que não há um método de imputação universal que se comporte como o melhor para todos os classificadores (LUENGO; GARCÍA; HERRERA, 2012), no entanto, é possível implementar estratégias que otimizem a acurácia de classificadores usando diferentes algoritmos, tal como considerado nos métodos propostos neste projeto de tese;

- Estudos comparativos com foco em domínios de aplicação onde a análise de séries temporais-espaciais é predominante demonstram que há uma lacuna na literatura quanto a utilização de métodos de imputação baseados em aprendizado de máquina.

4.4 MÉTODOS DE IMPUTAÇÃO BIOINSPIRADOS

A utilização de modelos bioinspirados no processo de imputação de dados ocorre de duas formas. Alguns trabalhos utilizam-no para otimizar parâmetros de outras técnicas - estas responsáveis por realizar a imputação; outros utilizam os modelos bioinspirados para buscar os valores a serem imputados. Dois trabalhos considerados precursores no uso de modelos bioinspirados no contexto da imputação de dados são Abdella e Marwala (2005a) e Abdella e Marwala (2005b), onde os autores utilizaram algoritmos genéticos para auxiliar no treino de uma rede neural artificial, a qual é responsável por predizer os valores a serem imputados.

Ainda na linha de trabalhos híbridos, de Andrade Silva e Hruschka (2009) propõem o algoritmo EACImpute que consiste em um algoritmo evolucionário para imputação baseada em agrupamento. A proposta deste método é utilizar a computação evolucionária para encontrar

grupos de instâncias semelhantes, nos quais um método de imputação baseado em vizinhos próximos é aplicado. Portanto, o modelo bioinspirado adotado é utilizado para determinar os grupos (*clusters*) onde aplica-se o KNNI. O EACImpute foi comparado com outros métodos de imputação de dados conhecidos, a saber, KNNI (TROYANSKAYA et al., 2001), SKNN (KIM; KIM; YI, 2004), IKNN (BRÁS; MENEZES, 2007), o KMI (HRUSCHKA; Hruschka Jr.; EBECKEN, 2005) e a imputação por média/moda. Os resultados mostraram que o EACImpute possui desempenho semelhante aos outros métodos analisados, portanto, passível de ser eleito para uso em aplicações reais. É interessante notar que todos os métodos analisados neste estudo (excetuando-se a imputação por média/moda) são baseados em vizinhança e agrupamento.

Ainda em trabalhos híbridos, França (2010) e Veroneze (2011b) propõem um método de imputação baseado em biclusterização por inteligência de enxame (SwarmBCluster) aplicado à identificação de bicluster, uma vez que os algoritmos tradicionais de biclusterização não podem ser aplicados em conjuntos de dados com VA; fato devido a medida utilizada para se avaliar a homogeneidade de um blicluster (erro médio residual), o qual não pode ser calculada corretamente, sendo necessária a utilização de algum método de tratamento prévio.

Por este motivo, os autores propuseram a seguinte estratégia: primeiro os valores ausentes são pré-imputados utilizando uma técnica tradicional. No trabalho de Veroneze (2011b) os autores utilizaram a imputação baseada em vizinhos próximos utilizando a distância euclidiana como medida de similaridade. Então aplica-se o SwarmBCluster a fim de identificar um conjunto de biclusters, em seguida, os valores pré-imputados são removidos. Utiliza-se então o erro médio residual e a coerência aditiva (duas medidas de desempenho do processo de biclusterização), calculados a partir do conjunto de dados pré-imputado para estimar os novos valores. Isto é feito assumindo que cada valor ausente de um bicluster é uma variável em um problema de programação quadrática, resolvendo-o, obtêm-se os valores a serem imputados.

Em (VERONEZE, 2011a) o autor analisa o método descrito em França (2010) e Veroneze (2011b), avaliando seu desempenho em relação ao custo computacional; quantidade de valores ausentes; e sensibilidade dos parâmetros como quantidade de formigas, erro mínimo aceitável na identificação dos biclusters e número de iterações. Como conclusão os autores indicam que o método proposto mostrou-se computacionalmente custoso em relação aos dois outros métodos analisados (KNNI e *regulated Singular Value Decomposition* (rSVD)), apesar de ter se mostrado competitivo. É importante notar que este método, além de requerer um método de imputação simples a fim de fornecer um "chute inicial" para o problema, requer também um método para resolver o problema de otimização quadrática resultante do processo. Ademais, os métodos de imputação utilizados não apresentaram bom desempenho no estudo comparativo conduzido por Luengo, García e Herrera (2012). Por fim, o trabalho de França, Coelho e Von Zuben (2013) apresenta-se como uma atualização do primeiro, havendo a retificação de pontos fracos da primeira abordagem, principalmente quanto a forma de como o modelo anterior estimava os valores ausentes. Logo, o mesmo também recai a uma análise em relação às mesmas

medidas e métodos sob o ponto de vista apenas de agrupamento dos dados.

Gautam e Ravi (2014) propõem um método de imputação baseado em clusterização evolutiva semelhante aos trabalhos supramencionados. O processo de imputação dá-se da seguinte forma, primeiro divide-se o conjunto de dados em duas partes (uma com os casos completos e outra com instâncias incompletas). Então aplica-se o algoritmo de evolução de grupos (*Evolving Clustering Method* (ECM) proposto por Song e Kasabov (2001) a fim de identificar todos os grupos. Para realizar a imputação de dados, calcula-se a distância euclidiana entre a instância com valor ausente e os centros dos grupos, sem levar em consideração os atributos com valor ausente. Identificado a qual grupo a instância com dados faltosos pertence, substitui-se o(s) atributo(s) com valor(es) ausente(s) pelos valores correspondentes ao centro do grupo. Portanto, o método evolucionário é usado neste trabalho apenas para a definição dos grupos, e não para a imputação propriamente dita. Adicionalmente, o *framework* experimental pode ser considerado incipiente, uma vez que o método proposto é comparado apenas contra a imputação por média e por outro método proposto pelo autor, o qual utiliza-se de K-Means e redes neurais artificiais; e os conjuntos de dados utilizados são compostos apenas de atributos numéricos.

Tang et al. (2015) propõem um método híbrido que também integra algoritmo genético à imputação por agrupamento usando *fuzzy C-means*. O AG é utilizado para otimizar dois parâmetros - a função de pertinência e os centróides do modelo *fuzzy* - pois a eficácia do método tradicional (sem o AG) é sensível à seleção dos valores iniciais o que resulta em soluções subótimas. O cálculo da função de aptidão é dada pelo erro entre a imputação e os valores atuais (imputados na etapa anterior), portanto, tal método continua sensível ao valor da primeira imputação.

Figueroa García, Kalenatic e López Bello (2008) propuseram um método de imputação para análise de séries temporais usando algoritmos evolucionários, mais especificamente, utilizando um algoritmo genético para minimizar a função de erro obtida a partir da função de autocorrelação, média e variância. Tais medidas foram escolhidas pois são estatísticas úteis para estimar a maior parte de modelos lineares (e.g. *Autoregressive Integrated Moving Average* (ARIMA) e *autoregressive Conditional Heteroskedasticity* (ARCH)) e são calculadas a partir dos casos completos, ao contrário dos trabalhos discutidos anteriormente (França (2010) e Veroneze (2011b)), onde os autores utilizavam-se do artifício de uma pré-imputação.

O algoritmo genético proposto por Figueroa García, Kalenatic e López Bello (2008) busca por valores a serem imputados de forma a minimizar a distância entre as características do conjunto de dados original (casos completos) e da base imputada. O tamanho do cromossomo é igual ao número de valores ausentes existentes no conjunto de dados e o número de indivíduos é definido por uma função que relaciona uma constante definida pelo usuário com o número de valores ausentes. Os autores também conduziram testes acerca da inicialização da população, tanto baseada na função densidade de probabilidade quanto na geração uniforme, sendo que a última apresentou melhores resultados, segundo os autores. Além do número de gerações como

critério de parada, adotou-se também a estratégia de um determinado número de gerações sem melhora na função de aptidão.

Em Figueroa García, Kalenatic e López Bello (2010) os autores extrapolam as análises presentes no trabalho anterior, acrescentando elitismo ao método proposto e avaliando estatisticamente o desempenho do método proposto. Em ambos os trabalhos, os autores utilizaram um conjunto de dados de séries temporais composto por 2734 exemplos, obtidos por duas capturas diárias nos 1367 dias de medição. Tais dados estavam dispostos em gráficos, sendo que VA eram identificados com a ausência de indicação escalar e que o autor não especificou uma medida importante relacionada à quantidade de valores ausentes, a taxa de instâncias com VA. A fim de estimar esta taxa, analisou-se cuidadosamente os gráficos apresentados e, estima-se que cerca de 65% das instâncias eram casos incompletos. Logo, apenas 35% das instâncias eram consideradas casos completos. Portanto, esta forma de modelagem da função de aptidão impõe um viés ao modelo, uma vez que a função de auto-correlação, a variância e a média são obtidas a partir de um subconjunto de dados pequeno e potencialmente não representativo. Apesar de não estar explícito no trabalho, infere-se que esta forma de modelagem assume que as variáveis obedecem o mecanismo de ausência MCAR.

Por fim, Figueroa García e colaboradores extrapolam o algoritmo genético descrito nos trabalhos anteriores para dados multivariados em Figueroa García, Kalenatic e López Bello (2011). Algumas modificações importantes no método são: i) algumas medidas são calculadas a partir de dados parcialmente completos - não removem-se todas as instâncias com VA, apenas as que possuem VA para o atributo de interesse; ii) a função de aptidão é modificada, não mais incluindo a variância em seu cálculo, apenas a média e a covariância.

Como conclusão dos trabalhos discutidos, os autores afirmam que a estratégia evolutiva é adequada e fácil de ser aplicada ao contexto de análise de séries temporais; sua flexibilidade e capacidade de aprender modelos não lineares a torna uma alternativa e uma ferramenta poderosa para gerar soluções apropriadas (bases imputadas); e que o método pode ser extrapolado para diferentes contextos por meio da modificação da função de aptidão e de alguns parâmetros do método. Tais considerações foram motivadoras para a utilização de estratégias evolucionárias para otimização no tratamento de dados ausentes por meio da imputação múltipla de dados.

Uma adaptação dos trabalhos de Figueroa Garcia é o de Krishna e Ravi (2013), onde os autores apresentam um método de imputação de dados baseado em otimização por enxame de partícula. As diferenças consistem em: i) a pré-imputação é realizada pela média ao invés do uso do KNNI; ii) a função de aptidão utilizando o erro quadrático médio entre as matrizes de covariância calculadas a partir dos casos completos e de todo o conjunto de dados incluindo as instâncias imputadas, e a diferença absoluta entre os determinantes das duas matrizes de covariância. Os autores, utilizam um *framework* experimental idêntico ao de Gautam e Ravi (2014), o qual dificulta o processo de comparação fidedigna, inclusive, apresentando resultados inferiores ao último autor.

Autores deste mesmo grupo de pesquisa propõem dois métodos de imputação híbridos, que combinam os métodos propostos anteriormente por eles, os quais baseiam-se em computação evolucionária, clusterização e redes neurais. O primeiro método proposto por (GAUTAM; RAVI, 2015) mescla o PSO utilizado por Krishna e Ravi (2013) com o ECM proposto por Gautam e Ravi (2014) e é denominado (PSO+ECM), sendo que a função do PSO é encontrar valores ótimos para parametrizar o ECM, o qual é responsável por realizar a imputação de dados na mesma lógica adotada em Gautam e Ravi (2014). Este novo método foi motivado pela sensibilidade da imputação baseada por ECM ao parâmetro *Dthr*, que é o valor limite para o radio de um cluster, o qual afeta sensivelmente a performance da imputação por evolução de grupos. Os demais parâmetros são idênticos à imputação por ECM puro, incluindo a função de aptidão que, como mencionado, é uma variante da função proposta por Figueroa García, Kalenatic e López Bello (2008).

No segundo método apresentado por (GAUTAM; RAVI, 2015), a imputação ocorre em duas fases. A primeira consiste em normalizar os dados no intervalo [0, 1] para então aplicar a imputação (PSO+ECM). A segunda fase consiste em aplicar o ECM no novo conjunto de dados gerado (com as instâncias imputadas), a fim de identificar os centros dos grupos, os quais serão utilizados como nós escondidos de uma rede neural auto-associativa treinada pelo algoritmo de aprendizado de máquina extremo a fim de predizer os atributos com VA nas instâncias pertencentes àquele grupo. É importante notar que este método requer uma fase de pré-processamento, onde os dados são normalizados entre o intervalo [0, 1], devido ao uso da rede neural.

Os autores avaliaram os métodos propostos da mesma forma dos trabalhos anteriores, utilizando 12 conjuntos de dados, contendo apenas atributos numéricos; os métodos considerados *baseline* são também propostos pelos autores em detrimento aos métodos bem-estabelecidos na literatura - fato que prejudica a comparação fidedigna com os demais métodos dispostos na literatura; e o desempenho computacional dos métodos propostos não são analisados. Todavia, devido a adição de múltiplos estágios e pela avaliação das técnicas utilizadas, acredita-se que estes métodos sejam computacionalmente custosos em relação ao estado-da-arte. Por fim, conforme discutido na Seção 4.3, a otimização de critérios baseados na covariância não necessariamente implica em uma melhora no critério baseado na acurácia do classificador (LIU; BROWN, 2013); e ainda, conforme Schafer e Graham (2002) e Hruschka et al. (2009), os métodos de imputação não podem ser devidamente avaliados aparte da tarefa de modelagem. Reitera-se que estas medidas são úteis na construção de modelos lineares para análise de séries temporais, por conseguinte, tais medidas podem ser consideradas potencialmente úteis para avaliação de métodos de imputação aplicados a este nicho.

Patil e Bichkar (2010) desenvolveram um algoritmo genético para tratar da problemática de dados faltosos sob o ponto de vista da imputação múltipla. Diferentemente dos trabalhos previamente analisados, os autores adotaram como função de aptidão a acurácia do classificador, conforme indicado na literatura. Entretanto, o estudo possui pontos divergentes dos demais

trabalhos na área, como uma não explicitação acerca do mecanismos de ausência das bases utilizadas; as medidas utilizadas também possuem pouca explicações de como foram mensuradas e tendem a confundir o leitor em relação às conclusões

Tran, Zhang e Andreae (2015) propõem um método de imputação múltipla baseado em programação genética como método de regressão para estimar os valores ausentes, chamado *Genetic Programming Multiple Imputation* (GPMI). Tal como em outros trabalhos discutidos, este método requer uma pré-imputação. No GPMI os autores escolhem um método de imputação aleatoriamente a partir de uma lista de sete algoritmos, incluindo a imputação por média, KNNI, *hot deck*, dentre outros. Então, aplicam-se os operadores genéticos, onde as variáveis terminais são os atributos sem valores ausentes e o atributo a ser imputado é a variável de interesse. Dois critérios de avaliação foram utilizados: a acurácia preditiva do método de imputação e a acurácia do classificador. Como conclusão os autores afirmam que o GMPI alcançou, na maior parte dos casos, resultados melhores quanto às medidas de desempenho utilizadas em comparação aos outros métodos, contudo, não apresentaram evidências estatísticas de tal fato. Como considerações deste método, têm-se: i) apesar de recentemente proposto, tal método restringe-se apenas a um tipo de atributo, o numérico; ii) a noção de evolução da(s) solução(ões) provenientes de outros métodos de imputação, as quais podem ser "evoluídas/combinadas" por métodos evolucionários a fim de produzir soluções mais robustas.

4.4.1 CONSIDERAÇÕES DE ANÁLISE

- O uso de estratégias evolucionárias na imputação de dados ocorre de duas formas: i) utilizando para de melhorar a convergência ou otimizar a parametrização de outros métodos, como na imputação por ECM e fuzzy c-means (de Andrade Silva; HRUSCHKA, 2009; FRANÇA, 2010; GAUTAM; RAVI, 2014; TANG et al., 2015) ; ii) realização da imputação propriamente dita (Figueroa García; KALENATIC; López Bello, 2008; Figueroa García; KALENATIC; López Bello, 2010; PATIL; BICHKAR, 2010; Figueroa García; KALENATIC; López Bello, 2011; KRISHNA; RAVI, 2013; TRAN; ZHANG; ANDREAE, 2015). Sendo a última o tema de interesse deste projeto de tese;

- A maior parte dos métodos que realizam a imputação por meio de modelos bioinspirados restringem-se a:

  - tratar apenas atributos de tipo numérico, vide (Figueroa García; KALENATIC; López Bello, 2008; Figueroa García; KALENATIC; López Bello, 2010; KRISHNA; RAVI, 2013; TRAN; ZHANG; ANDREAE, 2015);

  - utilizar estatísticas (e.g. covariância, média e variância) como heurísticas para guiar o processo de busca, tais medidas são úteis para construção de modelos lineares para análise de séries temporais, mas conforme Liu e Brown (2013), a otimização de critérios baseados na covariância não necessariamente implica em uma melhora

no critério baseado na acurácia do classificador. Portanto, a adoção unicamente de tais medidas não é indicada quando pretende-se otimizar o processo de classificação;

- Nos trabalhos analisados, os algoritmos genéticos foram mais presentes e apresentaram resultados superiores a outras heurísticas como o PSO, sendo este um dos motivos da adoção de estratégias evolucionárias neste projeto de tese;
- É possível combinar soluções obtidas por meio da aplicação de outros métodos de imputação por meio de uma estratégia evolucionária, obtendo resultados mais robustos e evitando soluções subótimas.

4.5 CONSIDERAÇÕES FINAIS

O tratamento de valores ausentes por meio da imputação de dados é um campo de estudo vasto, neste capítulo abordaram-se alguns trabalhos recentes que discorrem sobre imputação de dados que abalizaram este projeto de tese. Inicialmente apresentaram-se revisões na literatura as quais expuseram as principais abordagens e direcionamentos de pesquisas. Em resumo, por meio da análise destes trabalhos foi possível perceber que a imputação de dados é uma das estratégias mais eficientes para tratamento de valores ausentes, no entanto, vislumbrou-se também uma baixa adoção desta abordagem pela dificuldade dos pesquisadores em aplicá-la (*e.g.* dificuldades de parametrização, falta de *know-how* na área *etc*).

No segundo momento, estudos comparativos foram analisados, os quais indicaram métodos para lidar com VA a serem adotados como *baseline*, descreveram *frameworks* experimentais, medidas de desempenho e testes estatísticos adotados. Tais estudos apontaram que medidas de avaliação dos métodos de imputação baseadas em covariância ou no erro entre o dado imputado e o dado real não necessariamente estão relacionadas com um aumento na acurácia do classificador; e destacaram também que taxas de valores ausentes consideradas baixas (1 - 5%) podem impactar sensivelmente na tarefa de classificação, reiterando a necessidade de tratamento prévio.

A terceira e última seção abordou os métodos de imputação que utilizam-se de modelos bioinspirados, o que pode ocorrer de duas formas: utilizando-os para otimizar a parametrização ou melhorar a convergência de outro método de imputação, geralmente baseado em agrupamento; ou na estimação dos valores a serem imputados por meio desta classe de algoritmos. Neste âmbito, perceberam-se algumas lacunas na literatura quanto à necessidade de: i) formalização da imputação de dados como um problema de otimização combinatorial; ii) incluir na modelagem estratégias para lidar atributos de tipo nominal ao invés de tratar exclusivamente atributos de tipo numérico; iii) incluir informações acerca da construção do modelo de classificação na heurística de busca; iv) extrapolar e testar métodos de imputação no cenário de classificação multirrótulo.

5 GAIMP: IMPUTAÇÃO MÚLTIPLA DE DADOS BASEADA EM ALGORITMOS GENÉTICOS

5.1 CONSIDERAÇÕES INICIAIS

É inquestionável a importância da análise de dados na Era da Informação. Mesmo assim, a maior parte dos métodos de análise, sejam eles estatísticos ou baseados em aprendizado de máquina, não são robustos a um problema ubíquo na área, os valores ausentes. Com o intuito mitigar seus efeitos danosos, diversos estudos têm sido conduzidos visando desenvolver estratégias para lidar com a ausência dos dados, sendo a imputação de dados uma das mais bem aceitas, tanto pela academia quanto pela indústria.

Assim como vários outros problemas encontrados na mineração de dados, a ImpD pode ser modelada como um problema de otimização combinatorial, onde o objetivo é encontrar uma combinação de valores a substituir os que estão em falta, de forma a reduzir o viés imposto. Neste contexto, algoritmos evolucionários têm sido aplicados com sucesso para resolver problemas de otimização. Não obstante o paradigma sobre qual o método de imputação de dados se baseia, algumas limitações podem ser identificadas, principalmente no tocante aos tipos de dados e a análise de caso completo.

Portanto, conforme discuto nos Capítulos 1 e 4, o desenvolvimento de métodos de imputação flexíveis quanto ao tipo de dados e que evitem a análise de caso completo representam desafios de pesquisa pertinentes à área de foco desta tese de doutorado. Tendo em vista preencher tais lacunas na literatura, desenvolveu-se um método de imputação de dados baseado em algoritmos genéticos para otimizar a tarefa de classificação, chamado de GAImp, o qual tem como metas: tratar conjuntos de dados com atributos de tipos mistos de forma satisfatória, considerando instâncias incompletas e levando em consideração informação da geração do modelo, mais especificamente, a acurácia do classificador.

Apesar do estudo de caso analisado ser a classificação de padrões, chama-se atenção para a flexibilidade da proposta, a qual pode ser adaptada a outras tarefas que envolvam análise multivariada, como regressão, classificação multirrótulo, agrupamento e análise de séries temporais; e também, é possível otimizar múltiplas medidas de desempenho, inclusive conflitantes, por meio da adoção de estratégias multiobjetivo, conforme será abordado no Capítulo 6.

Neste capítulo, o método GAImp é apresentado, destacando as motivações da adoção do algoritmo genético como heurística de busca e descrevendo sua estrutura e fluxo de trabalho. Também apresenta-se o *framework* experimental adotado, que consiste em uma adaptação de Luengo, García e Herrera (2012), o qual levou em consideração seis classificadores, representando três grupos de métodos de classificação: indução de regras, modelos de aproximação e

aprendizado baseado em instâncias; sete métodos de imputação bem estabelecidos; e 15 conjuntos de dados que já possuem valores ausentes, obtidos a partir do repositório de aprendizado de máquina da Universidade da California, Irvine (LICHMAN, 2013). Por fim, os resultados para os cenários analisados são discutidos e as considerações finais do capítulo são apresentadas.

5.2 IMPUTAÇÃO MÚLTIPLA DE DADOS E ALGORITMOS EVOLUCIONÁRIOS

A imputação múltipla é uma das categorias de método de imputação que possuem mais vantagens, incluindo: a produção de estimativas imparciais, proporcionando maior robustez do que abordagens *ad hoc*; utilização de todos os dados disponíveis, preservando o tamanho da amostra; e ainda ser possível a utilização do software estatístico (ou de aprendizado de máquina) já utilizado pelos analistas (MCCLEARY, 2002). Tais vantagens são decorrentes da sua modularidade e estratégia de estimação dos valores ausentes e combinação das diversas soluções produzidas, como será visto adiante.

Por meio da análise da imputação múltipla é possível perceber algumas semelhanças/aplicabilidade dos algoritmos evolucionários. Portanto, esta seção destina-se a apresentar alguns conceitos de imputação múltipla, a fim de que seja estabelecido um paralelo/relação com os algoritmos evolucionários, para então descrever os métodos propostos adiante.

5.2.1 CONCEITOS DE IMPUTAÇÃO MÚLTIPLA

A imputação múltipla de dados foi proposta por Rubin (1987) e emergiu como uma alternativa flexível aos métodos baseados em verossimilhança para uma ampla variedade de problemas. A MI pode ser descrita como um processo de três etapas, as quais encontram-se descritas a seguir:

1. Imputação: cada valor ausente é substituído por uma lista de $m > 1$ valores plausíveis, substituindo o j-ésimo elemento de cada lista para cada valor ausente correspondente, $j = 1, \ldots, m$, são produzidos m versões alternativas dos dados completos (SCHAFER; GRAHAM, 2002);

2. Análise: cada conjunto de dados gerado pela etapa anterior é analisado da mesma forma por um determinado método, geralmente, pelo método que seria aplicado se os dados estivessem completos;

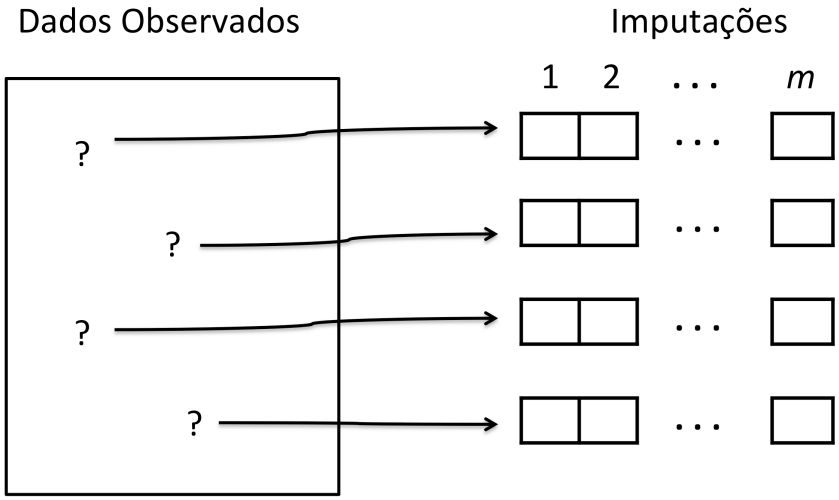
3. Combinação: os resultados obtidos na etapa anterior então são combinados a fim de obter uma solução que reflita as incertezas acerca de qual o melhor valor a ser imputado.

A Figura 13 descreve a etapa de imputação. Neste esquema, m conjuntos de dados imputados são gerados e são idênticos quanto à não existência de instâncias com valores ausentes,

mas diferem quanto aos valores imputados. A magnitude destas diferenças refletem as incertezas acerca de qual valor a ser imputado e nesta propriedade reside o poder deste método (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

E ainda, de acordo com Little e Rubin (2002, p. 85, tradução nossa), "*a desvantagem dos métodos de imputação simples é que a imputação de um único valor trata aquele valor como conhecido, e portanto, sem ajustes especiais, a imputação simples não consegue refletir a variabilidade amostral sobre um modelo de não resposta ou incerteza acerca do modelo correto para a não resposta*".

Figura 13 – Representação esquemática da imputação múltipla, onde *m* é o número de imputações.



Fonte: Elaborada pelo autor.

Historicamente, a segunda etapa é relacionada com a obtenção de alguma estatística de interesse, como média, coeficiente de regressão, coeficiente de correlação linear; por este motivo que medidas baseadas na covariância são tão difundidas. Conforme discutido no Capítulo 4, é importante que nesta etapa (análise), sejam levadas em consideração informações acerca da construção do modelo ou da tarefa de análise em questão, por conseguinte, o método proposto incorpora na fase de análise a acurácia dos classificadores.

Rubin (1987) desenvolveu um conjunto de regras para combinar as estimativas separadas e os erros padrão de cada um dos *m* conjuntos de dados imputados em uma estimativa geral, com um erro global, intervalos de confiança e *p* valores. Estas regras baseiam-se na teoria assintótica de verossimilhança da distribuição normal (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

Há também outros métodos para combinar as soluções, como por exemplo o uso da distribuição *posteriori* a partir de um subconjunto dos dados; a substituição dos dados de forma a criar padrões monotônicos (Ver Figura 3b); refinando aproximações sucessivas usando amostragem *etc* (LITTLE; RUBIN, 2002, p. 214).

Por fim, Rubin (1987) afirma que não é necessário um grande número de repetições para uma estimação precisa. A eficiência da imputação múltipla pode ser mensurada pela relação do parâmetro m com a taxa de ausência de dados de acordo com a Eq. 5.1.

$$Eficiência = (1+\lambda/m)^{-1} \tag{5.1}$$

Onde λ é a taxa de valor ausente. Por exemplo, se 50% das informações estão ausentes, 10 imputações ($m = 10$) é $100/(1+0,5) = 95\%$ de eficiência.

5.2.2 IMPUTAÇÃO MÚLTIPLA E ALGORITMOS EVOLUCIONÁRIOS

Conforme discutido no Capítulo 4, diversos modelos bioinspirados vêm sendo aplicados no contexto da imputação de dados, muitos dos quais utilizam-se de algoritmos evolucionários. As justificativas para este fato são as mais diversas e representam agentes motivadores para a adoção de algoritmos genéticos no método proposto, as seguintes merecem destaque:

Adaptabilidade à imputação múltipla: analisando os módulos e processos da imputação múltipla, é possível vislumbrar a aplicação de algoritmos evolucionários devido algumas semelhanças. Primeiramente, cada um dos m conjuntos de dados imputados pode ser visto como um indivíduo. Em seguida, a etapa de avaliação está intimamente relacionada com o cálculo de aptidão dos indivíduos da população. Já a combinação das soluções da imputação múltipla possui seu paralelo correspondente aos algoritmos evolucionários nos operadores de cruzamento e mutação. Por fim, múltiplas gerações ainda conferem a propriedade iterativa de acordo com a classificação dos métodos de imputação proposta por Zhang (2010), pois as soluções são sucessivamente refinadas por meio de um mecanismo de geração-e-teste, provendo maior confiabilidade ao resultado final. Tal paralelo não é encontrado com tanta clareza em outros modelos bioinspirados como na otimização por enxame de partícula e colônia de formigas;

Facilidade de codificação e paralelismo: a modularidade e simplicidade de seus operadores (*e.g* seleção, cruzamento, mutação) tornam o método fácil de ser codificado e independente do domínio de aplicação, excetuando-se o cálculo da função de aptidão. Ademais, os algoritmos genéticos possuem capacidade de paralelização implícita, decorrente da avaliação independente de cada indivíduo. Além de haver modelos de paralelismo bem estabelecidos (*e.g.* mestre e escravo, baseado em ilhas e modelos celulares), o que também favorece a adoção destes algoritmos em domínios de aplicação computacionalmente custosos;

Flexibilidade: a modularidade dos algoritmos genéticos, devido à separação entre o mecanismo de evolução e a representação particular do problema considerado, permite sua

fácil adaptação em diversos problemas. Isto é particularmente interessante para este projeto de tese, pois lhe estende um caráter de generalização, permitindo a investigação dos métodos propostos em outros domínios de estudo sem implicar em mudanças drásticas na proposta. É possível também a incorporação de conhecimento de fundo ou específico do problema, tanto na codificação e inicialização dos indivíduos, como será abordado no Capítulo 7, quanto na combinação e avaliação das soluções, conforme será abordado a seguir;

Auxílio na compreensão do problema: a análise da parametrização, seja por sintonia ou controle de parâmetros, e da convergência, permite uma melhor compreensão sobre o problema estudado, conforme será visto neste Capítulo e no próximo.

Frente aos motivos expostos, os algoritmos genéticos apresentam-se como uma alternativa interessante para implementar a imputação múltipla de dados no contexto multivariado e por isso foram escolhidos para compor o método proposto, o qual é descrito a seguir.

5.3 MÉTODO PROPOSTO: GAIMP

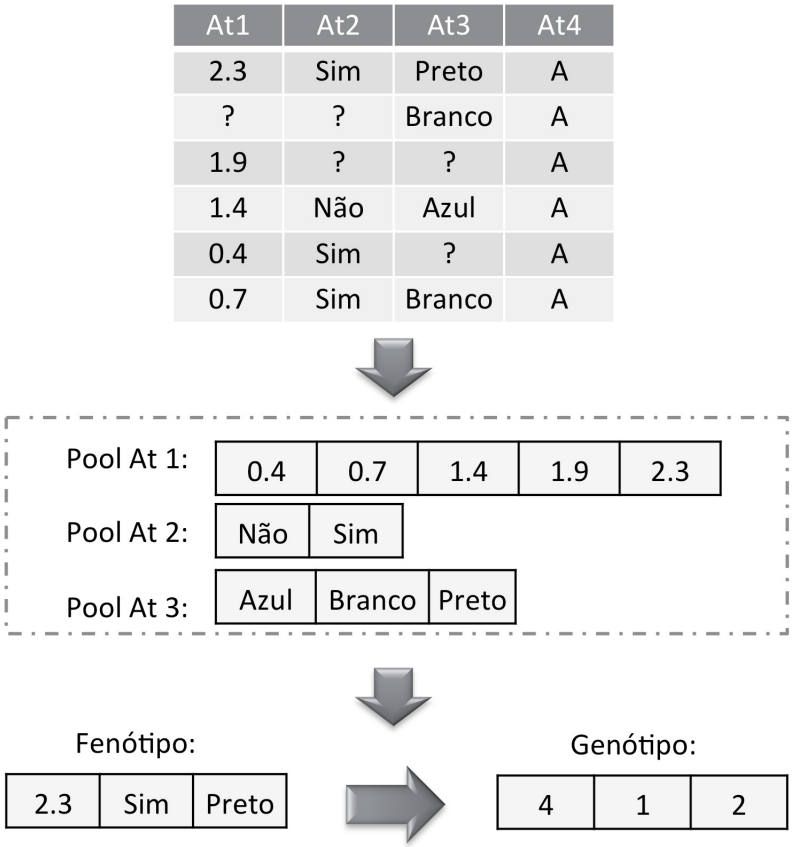
O método proposto obedece o raciocínio da imputação múltipla, onde geram-se m soluções, as quais são combinadas; além de também ser iterativo, pois as gerações vão refinando sucessivamente as combinações geradas. A representação e o mecanismo de combinação das soluções baseiam-se na ideia de subconjunto a fim de reduzir o espaço de busca, conforme discutido no Capítulo 3, o que consequentemente melhora a convergência do método. Antes de apresentar o fluxo de trabalho do método proposto, é necessário descrever o esquema de codificação do indivíduo e as estruturas de dados utilizadas.

5.3.1 CODIFICAÇÃO DO INDIVÍDUO

Como dito, o sistema de codificação do indivíduo é baseado na noção de subconjunto, portanto, o primeiro passo do GAImp é particionar o conjunto de dados, o qual é dividido em $x = q \cdot C$ subconjuntos de dados, onde q é uma constante definida pelo usuário e C é o número de classes. Esta estratégia reduz o espaço de busca $\mathbf{S}$, melhorando o tempo de convergência e permitindo a aplicação do método proposto em grandes conjuntos de dados.

Cada subconjunto de dados que possui valores ausentes é representado por um gene que irá compor o cromossomo. O gene consiste em alelos, os quais contém os valores a substituírem os ausentes para cada atributo. A Figura 14 ilustra a noção de gene adotado aqui. É importante salientar que os subconjuntos de dados são definidos em tempo de execução e permanecem o mesmo por toda a iteração, logo os indivíduos trabalham nos mesmos subconjuntos.

Figura 14 – Esquema de representação do gene.



Fonte: Elaborada pelo autor.

Dois outros conceitos da abordagem proposta são resumidos na Figura 14, o *pool* de soluções e a composição dos genes. O primeiro é responsável por conferir a capacidade do GAImp de lidar com atributos de tipos mistos, e consiste em uma estrutura de dados que ordena os conjuntos de todos os valores possíveis para cada atributo que possui valores ausentes em cada subconjunto de dados. Os gnes são construídos com base nessa estrutura.

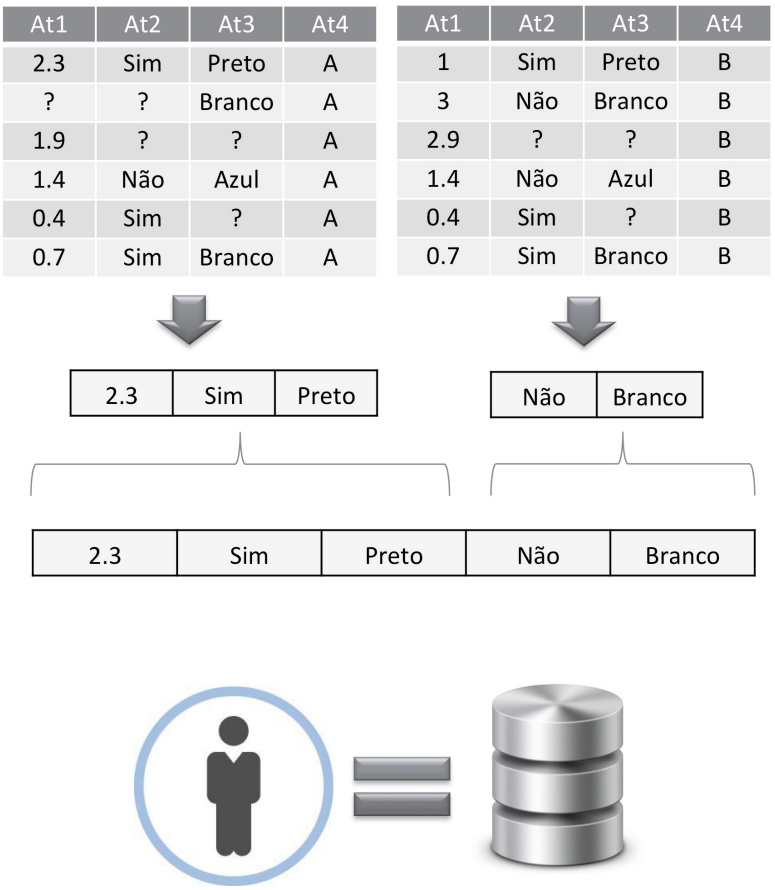
Chama-se atenção que, por meio de pequenas modificações no *pool* de soluções, é possível incluir conhecimento de fundo, como por exemplo, adotar uma estrutura de dados to tipo *HashMap <Chave, Quantidade>*, onde a "Chave" é o valor observado no conjunto (ou subconjunto) de dados e a "Quantidade" representa o número de vezes que o valor é observado, e que pode ser usado para dar maior probabilidade de imputação aos valores mais frequentes; é possível também substituir valores únicos de atributos numéricos por faixas de valores ou valores médios, reduzindo ainda mais o espaço de busca; dentre outros.

O segundo conceito contido na Figura 14 é o de genótipo, o qual é formado pelo índice do *pool* de soluções. O mapeamento genótipo fenótipo é realizado consultando o valor do índice correspondente. Utilizando o exemplo ilustrativo da Figura 14, tem-se três atributos com valores ausentes (At1, At2 e At3), consequentemente produz-se três *pools* de soluções (um para cada atributo) contendo todos os valores possíveis. No exemplo dado, o genótipo é composto

pelos índices do *pool* {4, 1, 2} e o fenótipo correspondente é {2.3, Sim, Preto}. O processo de imputação dá-se pela substituição do valor do fenótipo no atributo correspondente, no caso, o At3 possui o valor correspondente "Preto" e após a imputação passará a ser: {*Preto, Branco, Preto, Azul, Preto, Branco*}, a imputação dos outros atributos seguem a mesma lógica.

Conforme visto, o gene representa a solução de um subconjunto com VA. A solução para todo o conjunto de dados é montada no cromossomo, o qual possui tamanho igual ao número de atributos com VA em cada subconjunto, uma vez que cada atributo é representado por um alelo, conforme disposto na Figura 15.

Figura 15 – Esquema de codificação do cromossomo e representação do indivíduo.



Fonte: Elaborada pelo autor.

O exemplo apresentado na Figura 15 mostra um conjunto de dados dividido em três subconjuntos. Apenas dois deles apresentam VA e neles aplicou-se o esquema de codificação do indivíduo. Como resultado, dois genes foram criados e montados para formar o cromossomo. À vista disso, os valores de cada gene substituem os ausentes para cada um dos atributos de seu respectivo subconjunto. Em resumo, o cromossomo é uma montagem de todos os genes, combinando as soluções de cada subconjunto e resultando em uma solução completa. Analisando o esquema de codificação do cromossomo e representação do indivíduo, é possível perceber que o GAImp trata atributos do tipo numérico ou categórico da mesma forma, e que as relações

entre genes e cromossomos (atributos) são exploradas por meio do processo evolucionário, o qual é guiado pela função de aptidão. Este fluxo de trabalho é melhor apresentado adiante.

5.3.2 FLUXO DE EXECUÇÃO

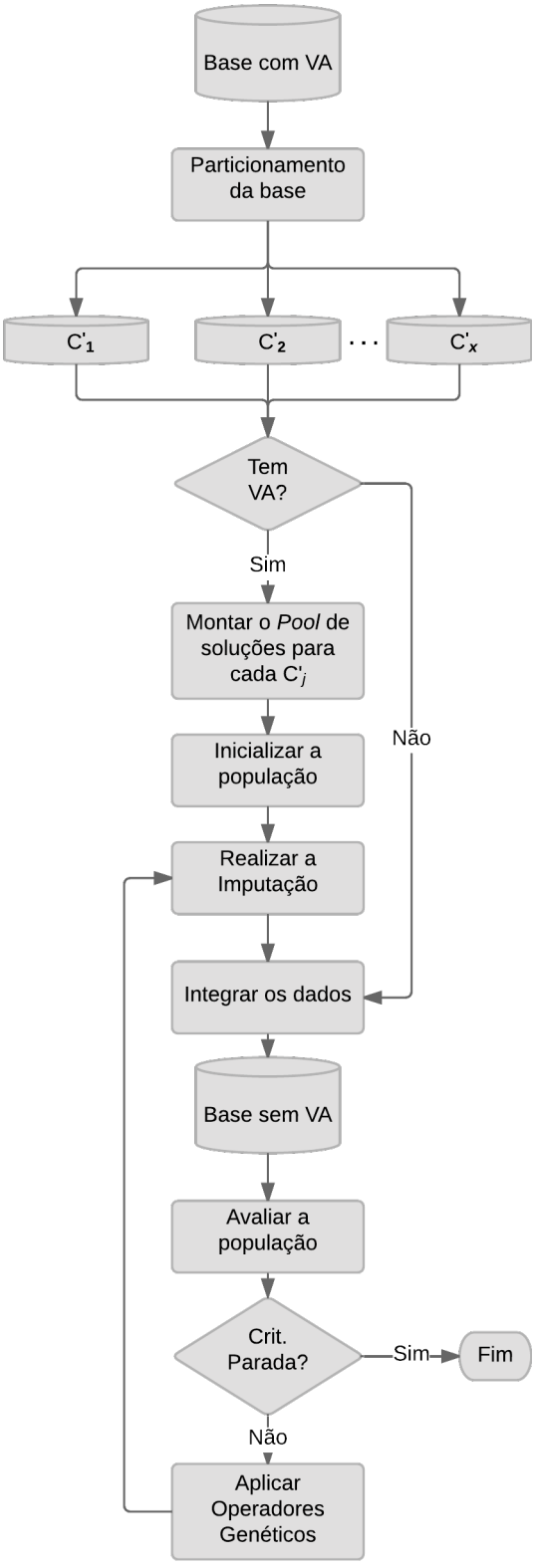
O processo evolucionário do GAImp segue as etapas básicas de um algoritmo genético, conforme segue:

1. **Início:** a população de indivíduos é inicializada, lembrando que cada indivíduo é uma solução $S \in \mathbf{S}$ e produzirá um conjunto de dados imputado;
2. **Avaliação:** a avaliação da função de aptidão (f) é feita para cada solução. Neste passo, o critério de parada também é avaliado, se não alcançado, então continua-se;
3. **Formação da nova geração:** novos indivíduos são criados por meio da aplicação dos seguintes passos:
    a) **Elitismo:** automaticamente passa-se um determinado número de indivíduos mais aptos para a próxima geração;
    b) **Seleção de pais:** selecionam-se dois parentes da população atual de acordo com suas aptidões, onde também deve-se garantir a manutenibilidade da diversidade;
    c) **Cruzamento:** promove a troca de material genético entre os indivíduos selecionados na etapa anterior, gerando dois novos indivíduos;
    d) **Mutação:** modifica randomicamente alguns valores dos cromossomos, provendo variabilidade genética aos novos indivíduos
4. **Aceitação:** promove-se os descendentes à nova população;
5. **Repetir:** retorna-se à etapa 2.

Estabelecendo um paralelo com a imputação múltipla, a Etapa 1 é a geração dos m conjuntos de dados imputados, sendo m o número indivíduos. A Etapa 2 é a subsequente avaliação das m soluções e as Etapas 3 e 4 representam a combinação das soluções, com o diferencial que o método proposto ainda inclui o passo iterativo, a refinar ainda mais as m soluções por meio da etapa 5. A Figura 16 resume o fluxo de trabalho do GAImp, explicitando os passos precedentes ao processo evolutivo e o encadeamento dos procedimentos.

Neste ponto, alguns detalhamentos do GAImp são necessários. O primeiro diz respeito ao processo de particionamento dos conjuntos de dados em x subconjuntos, onde optou-se por realizá-lo de forma aleatória, provendo independência em relação à classe e proporcionando variabilidade às soluções; ressalta-se que devido à flexibilidade do GAImp é possível incluir uma etapa de agrupamento, tal como adotado em métodos como o EACImpute e a imputação baseada em ECM (de Andrade Silva; HRUSCHKA, 2009; GAUTAM; RAVI, 2014).

Figura 16 – Fluxo de trabalho do GAImp.



Fonte: Elaborada pelo autor.

Para a construção do *pool* de soluções, duas abordagens foram testadas. Em uma, o *pool* de soluções para cada atributo com valor ausente foi construído a partir do conjunto de dados

completo (incluindo instâncias com valores ausentes), e na outra, a estrutura foi construída a partir das subpartições. Preferiu-se a última devido à decisão anterior, de se particionar a fim de reduzir o espaço de busca.

A integração consiste em agregar os subconjuntos de dados imputados àqueles sem valores ausentes, a fim de formar o conjunto de dados imputado correspondente. Cada indivíduo resulta em um conjunto de dados sem valores ausentes que é então avaliado. Se o critério de parada é atendido, o algoritmo seleciona o melhor indivíduo que representará a solução final do algoritmo, senão, os operadores genéticos são aplicados a fim de combinar as soluções e gerar a nova população, continuando o processo iterativo da imputação múltipla. A próxima Seção descreve os operadores genéticos e a função de aptidão usados no GAImp.

5.3.3 OPERADORES GENÉTICOS E FUNÇÃO DE APTIDÃO

Diversos operadores genéticos foram utilizados na fase de concepção do GAImp, a decisão final de quais operadores utilizar ocorreu após testes controlados avaliando-se os impactos no resultado final e a convergência do método. O método para seleção de parentes por torneio foi escolhido, pois, ao mesmo tempo que beneficia-se indivíduos mais aptos, a estocasticidade do método provê a variabilidade desejada ao GAImp.

O elitismo foi adotado a fim de que as melhores soluções não sejam perdidas durante o processo evolutivo. Como operador de cruzamento, uma variante do cruzamento de múltiplos pontos foi adotada, onde o número de pontos é igual ao número de genes, desta maneira, a informação genética é trocada por gene (subconjunto de dados), gerando novos indivíduos.

O mecanismo de mutação depende do tipo de atributo. Para atributos numéricos, a mutação Gaussiana (*creep mutation*) é adotada, onde assume-se o valor atual do índice como a média, então um número aleatório (e factível) é gerado de acordo com a distribuição normal. Como os valores estão ordenados no *pool* de soluções, esta estratégia de mutação possui menor probabilidade de causar mudanças abruptas nos valores. Já para os atributos categóricos, a mutação é realizada de forma completamente aleatória, selecionando-se um valor randômico a partir do *pool* de soluções para aquele atributo.

Como critério de parada optou-se pelo número de gerações, devido a sua simplicidade na calibração, levando-se em consideração algumas características do conjunto de dados importantes ao domínio de aplicação, como a dimensionalidade e a quantidade de valores ausentes do conjunto de dados.

Por fim, o método proposto foi projetado tencionando considerar informações contidas nas instâncias incompletas e tratar conjuntos de dados com tipos de atributos mistos, o que é realizado por mérito do *pool* de soluções e da função de aptidão. Esta última busca levar em consideração informações acerca da construção do modelo, por este motivo, a medida de desempenho de uma solução candidata à imputação final é calculada de acordo com a média da

acurácia de classificadores, representando três grupos de métodos de classificação: indução de regras, modelos de aproximação e aprendizado baseado em instâncias. O cálculo da função de aptidão é feito de acordo com a Eq. 5.2.

$$Fitness = \frac{1}{l}\sum_{i=1}^{n} acc_i \tag{5.2}$$

Na Eq. 5.2, l é o número de classificadores usados e acc_i é a acurácia do i-ésimo classificador. O método proposto não define quais classificadores devem ser usados, apenas indica a utilização de métodos que representem diferentes paradigmas de aprendizado. Apesar de outras formas de agregação serem passíveis de adoção (*e.g.* mediana, média harmônica), a média foi a que apresentou melhores resultados. Fato também observado por Barros, Basgalupp e Carvalho (2015), uma vez que a agregação da acurácia pela média foi a que obteve melhores resultados nos cenários analisados (em relação à acurácia e a despeito de medida F1).

Evidencia-se também que a modularidade do método proposto permite, sem que se altere o fluxo de trabalho ou os demais operadores genéticos, a adaptação para outros domínios de aplicação por meio de uma modificação no cálculo da função de aptidão, como por exemplo, adotar o erro quadrático médio para a tarefa de regressão; medidas baseadas na covariância para a análise de séries temporais (Figueroa García; KALENATIC; López Bello, 2010); ou até mesmo múltiplas medidas de avaliação (em uma abordagem multiobjetivo) como a medida F1 e o *Exact Match* (EM) no caso da classificação multirrótulo ou da classificação de conjuntos de dados desbalanceados (BARROS; BASGALUPP; CARVALHO, 2015; GONCALVES; PLASTINO; FREITAS, 2013). Apesar da variedade de aplicações possíveis, escolheu-se a classificação como estudo de caso principal para se avaliar o desempenho do método proposto, conforme é abordado a seguir.

5.4 EXPERIMENTOS COMPUTACIONAIS

Os experimentos realizados para avaliar o desempenho do método proposto foram conduzidos de forma semelhante ao *framework* experimental adotado por Luengo, García e Herrera (2012). Esta escolha deu-se devido à relevância e robustez das análises realizadas no referido estudo, conforme discutido no Capítulo 4.

Testes exaustivos foram executados usando um servidor com 12 processadores de 2,1 Ghz cada, 16 Gigabytes de memória RAM, rodando a distribuição Linux CentOS 6.5 e Máquina Virtual Java versão 7, *update* 65. A fim de reduzir o gargalo de entrada e saída, os arquivos referentes aos conjuntos de dados foram lidos/escritos em uma partição virtual em memória RAM, usada como disco rígido. Nos referidos testes, buscou-se avaliar não somente a qualidade das soluções obtidas, mas também a influência dos parâmetros, convergência do método e custos computacionais, conforme é visto adiante.

5.4.1 METODOLOGIA EXPERIMENTAL

Os testes foram conduzidos usando seis métodos de classificação: C4.5 e PART representando o aprendizado por indução de regras; Máquina de Vetor de Suporte (*nu-Support Vector Machine* (nu-SVM)) e Naïve-Bayes, os quais pertencem a categoria de modelos de aproximação; e como representantes do aprendizado baseado em instâncias têm-se o KNN e a aprendizagem localmente ponderada (*Locally Weighted Learning* (LWL)). A Tabela 1 apresenta os parâmetros padrões utilizados.

Tabela 1 – Parâmetros dos classificadores.

Classificadores	Parâmetros
C4.5	Poda = sim, confiança = 0.25, instâncias por folhas = 2
PART	confiança = 0.25, itens por folha = 2
nu-SVM	Kernel = poly., nu = 0.1, eps = 0.001, passo = 1, gamma = 0.01, coef 0 = 0, p = 1, retroceder = sim
Naïve-Bayes	sem parâmetro
K-NN	K = 3, função de distância = Euclideana
LWL	K = 3, função kernel = constante

Os conjuntos de dados utilizados nos experimentos já possuíam valores ausentes e que encontram-se disponibilizados no repositório da Universidade da Califórnia, Irvine (LICHMAN, 2013). Estes conjuntos de dados foram escolhidos a fim de proverem diversidade ao processo de análise, pois incluem bases com atributos com tipo exclusivamente numéricos ou categóricos e com atributos com tipos mistos; diferentes taxas de VA - tanto quantidade geral quanto número de instâncias; diferentes dimensionalidades e quantidade de rótulos; dentre outros. A Tabela 2 apresenta a descrição resumida dos conjuntos de dados selecionados. Para treinar os classificadores, utilizou-se o método validação cruzada *k-fold* com 10 subconjuntos.

Para fins de melhorar a legibilidade da apresentação dos resultados optou-se por usar os acrônicos dos nomes dos conjuntos de dados. As demais informações da Tabela 2 são “At.”,’ “Cat.”, “Num.” “Inst.” que denotam quantidades de: atributos, atributos de tipo categórico, atributos de tipo numérico, e instâncias. Os demais acrônicos que possuem o sufixo VA, dizem respeito às respectivas quantidades com valor ausente, por exemplo “At. VA” indica a quantidade de atributos na base de dados a apresentar algum valor ausente em seus registros.

O desempenho do GAImp foi comparado contra oito métodos para lidar com valores ausentes, a saber: Imputação por Média ou Moda (MC); Concept most common attribute value for symbolic attribute and concept average value for numeric attribute; k-Nearest Neighbor imputation; Weighted Imputation with k-Nearest Neighbor (WKNNI); Ignore Missing (IM); Event Covering (EC); k-means Clustering Imputation (KMI). Para os métodos que requeriam parametrização, utilizou-se os valores indicados pelos autores dos mesmos, conforme disposto na Tabela 3.

Tabela 2 – Descrição dos conjuntos de dados usados nos experimentos.

Nome da base	*Acrônimo*	*At.*	*At. VA*	*Cat.*	*Cat. VA*	*Num.*	*Num. VA*	*Inst.*	*Inst. VA*	*Inst. VA(%)*	*VA(%)*
audiology	AUD	70	7	70	7	0	0	226	222	98,23	2,00
autos	AUT	26	7	11	1	15	6	205	46	22,44	1,10
bands	BAN	20	19	1	0	19	19	539	174	32,28	5,11
breast	BRE	10	2	10	2	0	0	286	9	3,15	0,31
cleveland	CLE	14	2	1	0	13	2	303	6	1,98	0,14
credit	CRT	16	7	10	5	6	2	690	37	5,36	0,60
dermatology	DER	35	1	1	0	34	1	366	8	2,19	0,06
hepatitis	HEP	20	15	14	10	6	5	155	75	48,39	5,38
horse-colic	HOC	23	21	16	14	7	7	368	361	98,10	22,76
house-votes-84	HOV	17	16	17	16	0	0	435	203	46,67	5,30
lung-cancer	LUN	57	2	57	2	0	0	32	5	15,63	0,27
mammographic	MAM	6	5	1	0	5	5	961	131	13,63	2,80
post-operative	POS	9	1	9	1	0	0	90	3	3,33	0,37
primary-tumor	PRT	18	5	18	5	0	0	339	207	61,06	3,68
wisconsin	WIS	10	1	1	0	9	1	699	16	2,29	0,22

Tabela 3 – Parâmetros dos métodos de imputação.

Método	Parâmetros
KNNI, WKNNI	K = 10
KMI	K = 10
	Erro = 100
	Iterações Máximas = 100
EC	T = 0.05

Como medidas de avaliação dos métodos de imputação, as acurácias dos classificadores e o Wilson's Noise Ratio (Eq. 4.5) foram adotadas. Ambos refletem a influência dos dados ausentes na classificação de padrões; no entanto, o WNR pode ser considerado o mais sensível pois ele avalia justamente a acurácia de classificação das instâncias com valores ausentes, possibilitando investigar o impacto do método da imputação diretamente (LUENGO; GARCÍA; HERRERA, 2012). Novamente, para facilitar as comparações, adotou-se a normalização; neste caso o WNR foi invertido, portanto, quanto maior o valor, melhor o desempenho do método.

Para mensurar a significância estatística obtida pelas comparações entre os métodos de imputação, o teste de Wilcoxon pareado com intervalo de confiança de 95% foi aplicado para as acurácias dos classificadores; pois ao passo de não ser paramétrico, é considerado seguro e robusto para comparações estatísticas pareadas (DERRAC et al., 2011); adicionalmente, as diferenças entre as acurácias obtidas são pequenas, reforçando a indicação deste teste estatístico,

sobretudo pela sua simplicidade. Para o WNR, como observou-se diferenças mais acentuadas e a medida é considerada crítica para avaliar o impacto da imputação, optou-se pelo teste de Friedman.

Para parametrização do método, diversos experimentos foram conduzidos visando analisar o impacto dos parâmetros quantitativos no desempenho, tanto em termos das medidas de desempenho adotadas (convergência e diversidade) quanto em relação ao custo computacional. A Tabela 4 apresenta os parâmetros adotados. Devido ao caráter estocástico dos algoritmos genéticos, os resultados apresentados a seguir consistem na média de cinco execuções independentes.

Tabela 4 – Parâmetros GAImput.

Parâmetros	Valor
Tamanho da população	50
Taxa de mutação	10%
Taxa de cruzamento	90%
Indivíduos elitistas	3
Indivíduos por torneio	4
Número de gerações	30
q	3

5.4.2 AVALIAÇÃO DE DESEMPENHO

Esta subseção apresenta os resultados obtidos considerando a metodologia experimental acima descrita, seguida de discussões acerca da influência dos parâmetros na convergência do método.

5.4.2.1 RESULTADOS PARA A ACURÁCIA

A Tabela 5 apresenta o desempenho de cada método de imputação a respeito da acurácia dos classificadores. Os melhores resultados para a combinação entre algoritmo de classificação e os conjuntos de dados estão destacados em negrito.

Tabela 5 – Desempenho de cada método de imputação em relação à acurácia dos classificadores.

	Bases	**Métodos de Tratamento de Valores Ausentes**							
		CMC	EC	IM	KMI	KNNI	MC	WKNNI	GAImp
C4.5	AUD	**79,646**	77,876	0	77,876	77,876	77,876	77,876	76,991
	AUT	**86,341**	83,902	77,358	80,976	80,488	79,024	78,537	77,561
	BAN	**74,026**	66,976	66,027	68,831	68,46	70,872	69,202	69,759
	BRE	75,524	74,126	74,007	74,126	75,175	75,524	75,175	**76,923**
	CLE	53,135	55,116	54,209	52,475	52,805	55,446	52,805	**55,776**
	CRT	85,507	**86,377**	85,299	85,507	85,507	85,942	85,507	86,232
	DER	**95,355**	**95,355**	**95,531**	**95,355**	**95,355**	**95,355**	**95,355**	**95,355**
	HEP	**87,097**	81,29	83,75	74,839	78,71	79,355	76,129	84,516
	HOC	95,38	84,511	**100**	83,967	84,511	83,696	84,511	84,239
	HOV	**97,011**	96,782	96,552	96,782	**97,011**	**97,011**	**97,011**	**97,011**
	LUN	37,5	37,5	**81,481**	37,5	37,5	37,5	37,5	40,625
	MAM	**84,183**	83,039	82,892	82,31	83,663	82,622	83,871	81,79
	POS	**70**	**70**	68,966	**70**	**70**	**70**	**70**	68,889
	PRT	**50,147**	48,083	40,909	42,478	43,068	42,183	43,068	47,493
	WIS	95,136	95,422	**95,461**	95,422	95,422	94,421	95,422	93,991
PART	AUD	**81,858**	78,761	0	77,876	78,319	76,106	78,319	79,646
	AUT	**79,512**	79,024	76,101	75,61	77,073	73,171	75,61	78,049
	BAN	**72,542**	67,161	67,123	70,13	62,523	68,275	67,904	70,872
	BRE	72,028	69,58	70,758	70,28	71,329	**72,028**	71,329	71,329
	CLE	52,805	49,835	49,158	51,815	49,505	**53,465**	49,505	52,805
	CRT	83,478	84,203	86,064	85,507	83,768	84,203	83,768	85,797
	DER	**95,628**	**95,628**	94,134	**95,628**	**95,628**	**95,628**	**95,628**	94,262
	HEP	**91,613**	82,581	83,75	81,935	76,129	80	78,065	85,161
	HOC	96,467	80,978	**100**	78,533	76,902	80,163	80,978	81,522
	HOV	96,782	96,322	95,69	94,943	97,011	95,632	97,011	**97,241**
	LUN	40,625	37,5	**44,444**	37,5	37,5	37,5	37,5	40,625
	MAM	83,247	82,622	82,651	81,478	82,622	81,582	**83,039**	82,31
	POS	58,889	58,889	57,471	58,889	58,889	58,889	58,889	**60**
	PRT	**49,558**	42,183	39,394	43,363	43,658	36,578	43,658	48,378
	WIS	94,993	**96,137**	95,461	95,279	93,848	95,279	93,848	94,134
muSVM	AUD	**47,788**	46,46	50	45,575	**47,788**	44,248	**47,788**	45,133
	AUT	34,634	**35,61**	33,962	34,634	34,634	34,634	34,634	33,659
	BAN	58,442	58,813	**62,74**	58,442	58,813	58,442	58,813	58,071
	BRE	70,629	70,28	**71,48**	70,28	70,629	70,629	70,629	69,93
	CLE	**54,125**	**54,125**	53,872	**54,125**	**54,125**	**54,125**	**54,125**	**54,125**
	CRT	56,377	56,522	**60,184**	55,942	56,522	55,797	56,232	56,232
	DER	93,443	93,716	93,296	93,716	93,443	93,716	93,443	**93,989**
	HEP	79,355	79,355	**83,75**	79,355	79,355	79,355	79,355	79,355
	HOC	74,728	72,011	**100**	69,837	72,826	70,652	71,739	73,913
	HOV	96,092	95,172	**96,983**	96,092	96,092	95,632	96,092	96,322
	LUN	**40,625**	**40,625**	37,037	**40,625**	**40,625**	**40,625**	**40,625**	**40,625**
	MAM	**81,374**	79,188	79,88	80,125	80,333	79,084	80,125	80,853
	POS	71,111	71,111	**71,264**	71,111	71,111	71,111	71,111	71,111
	PRT	**50,442**	45,428	40,909	47,788	46,018	41,003	46,018	49,263
	WIS	95,994	95,994	**96,193**	95,994	95,851	95,994	95,851	96,137

Tabela 5 – Continuação.

	Bases	Métodos de Tratamento de Valores Ausentes							
		CMC	EC	IM	KMI	KNNI	MC	WKNNI	GAImp
NAIVE	AUD	**74,336**	69,912	50	69,912	70,354	72,124	70,354	73,451
	AUT	56,585	56,585	**57,862**	53,659	53,659	54,634	53,659	56,585
	BAN	65,121	63,822	64,658	63,636	64,75	65,677	64,935	**71,243**
	BRE	72,727	73,077	**75,09**	72,378	72,727	72,727	72,727	73,427
	CLE	55,446	55,116	**55,892**	55,446	54,785	54,455	54,785	55,776
	CRT	**77,971**	77,826	77,948	**77,971**	77,826	**77,971**	77,826	77,826
	DER	97,541	97,541	97,207	97,541	97,541	97,541	97,541	**97,814**
	HEP	**86,452**	83,226	82,5	84,516	85,161	83,226	85,161	85,806
	HOC	**87,772**	76,087	100	77,446	77,989	76,087	77,446	80,163
	HOV	91,494	90,345	91,379	**90,575**	91,034	90,345	91,034	90,805
	LUN	53,125	53,125	**62,963**	53,125	50	50	50	46,875
	MAM	**82,622**	81,27	81,205	80,957	81,27	81,374	81,27	82,518
	POS	65,556	65,556	66,667	65,556	65,556	65,556	65,556	**68,889**
	PRT	**54,277**	50,147	46,97	48,673	49,558	46,903	49,558	51,917
	WIS	95,994	96,137	**96,34**	96,137	95,994	96,137	95,994	95,708
LWL	AUD	48,23	**51,327**	50	**51,327**	50,885	49,115	50,885	46,903
	AUT	50,244	51,22	**51,572**	50,732	50,732	50,244	50,732	48,293
	BAN	**67,347**	60,111	63,014	64,935	59,74	62,338	59,74	67,161
	BRE	72,727	**73,077**	72,202	72,727	72,378	72,727	72,378	**73,077**
	CLE	57,756	57,096	56,229	**56,766**	57,426	57,756	57,756	56,766
	CRT	85,507	85,507	**86,371**	85,507	85,507	85,507	85,507	85,507
	DER	82,514	82,514	**82,961**	82,514	82,514	82,514	82,514	81,421
	HEP	**89,032**	80,645	86,25	81,29	78,065	76,129	76,129	**89,032**
	HOC	91,576	81,522	**100**	81,522	81,522	81,522	81,522	81,25
	HOV	96,322	95,402	**96,983**	96,092	96,092	95,632	96,092	96,322
	LUN	**53,125**	50	51,852	**53,125**	50	50	50	40,625
	MAM	81,998	81,79	**82,771**	81,998	81,894	81,894	81,894	81,998
	POS	68,889	68,889	**70,115**	68,889	68,889	68,889	68,889	68,889
	PRT	**42,773**	36,873	37,879	38,643	39,823	36,873	39,823	40,118
	WIS	91,273	91,416	91,947	91,416	91,416	91,702	91,273	**92,275**
3NN	AUD	**73,451**	68,142	0	69,027	67,699	65,929	67,699	73,009
	AUT	66,829	68,293	**70,44**	66,829	66,829	65,366	67,317	69,268
	BAN	69,573	70,686	69,315	69,202	68,275	70,315	69,202	**76,067**
	BRE	73,776	73,077	**76,173**	73,427	73,776	73,776	73,776	73,776
	CLE	52,805	52,805	57,239	53,135	53,135	52,475	53,135	**57,426**
	CRT	**86,087**	85,942	84,533	85,942	86,087	86,087	86,087	86,522
	DER	96,995	96,995	96,927	96,995	96,995	96,995	96,995	**97,268**
	HEP	83,871	85,161	83,75	80,645	79,355	81,29	79,355	**86,452**
	HOC	89,402	76,902	**100**	80,163	80,435	79,076	79,891	85,598
	HOV	94,483	93,563	92,241	93,563	94,253	93,563	94,253	**95,172**
	LUN	43,75	40,625	**48,148**	40,625	46,875	46,875	46,875	46,875
	MAM	79,813	78,98	78,193	78,876	78,876	79,188	79,501	**81,79**
	POS	**71,111**	**71,111**	67,816	**71,111**	**71,111**	**71,111**	**71,111**	70
	PRT	49,558	44,838	43,182	46,313	40,708	38,938	40,708	**52,212**
	WIS	96,423	96,423	96,779	96,423	96,423	96,567	96,423	**96,996**

Conforme pode ser visto na Tabela 5, há uma variação do desempenho dos métodos de imputação em relação aos classificadores, por exemplo, o GAImp obteve a melhor acurácia em apenas duas das 15 bases para o classificador PART, em contrapartida, o mesmo método

destacou-se no classificador KNN onde obteve melhores acurácias em 8 dos 15 conjuntos de dados.

Também é possível perceber que os valores das acurácias têm pouca ou nenhuma variação para algumas bases, como por exemplo em *dermatology*, *house-votes-84* e *post-operative*. Por este motivo, a avaliação de desempenho dos métodos foi realizada aplicando-se o teste pareado de Wilcoxon seguindo a mesma abordagem proposta por Luengo, García e Herrera (2012), onde obteve-se o ranqueamento dos métodos de imputação para cada classificador, conforme apresentado na Tabela 6.

Tabela 6 – Teste pareado de Wilcoxon aplicado a todos classificadores.

	C4.5	PART	nuSVM	NAIVE	LWL	3NN	Avg,	**Rank**
CMC	1,5	1	1	1,5	1	3	1,5	1
EC	6	6	3,5	6	6	4,5	5,333333333	4,5
IM	3	3	3,5	3	2	2	2,75	2,5
KMI	6	6	6,5	6	3	7	5,75	6,5
KNNI	6	6	3,5	6	6	7	5,75	6,5
MC	6	6	8	6	8	7	6,833333333	8
WKNNI	6	6	3,5	6	6	4,5	5,333333333	4,5
GAImp	1,5	2	6,5	1,5	4	1	2,75	2,5

Analisando o resultado apresentado na Tabela 6, é possível afirmar que o GAImp supera os demais métodos de imputação, mesmo com a perda para o CMC e o empate com o IM. Isto pode ser dito por dois motivos principais: i) o método CMC utiliza o conceito de subpartições considerando a classe como único critério de divisão, deste modo, os atributos com dados faltantes em cada participação terão seus VA substituídos pela média e moda da partição para atributos numéricos e categóricos, respectivamente, produzindo resultados enviesados como consequência; ii) o IM é considerado uma abordagem ingênua e que provoca perda de informação potencialmente útil, uma vez que as instâncias com valores ausentes são removidas do conjunto de dados. Logo em seguida, empatados, estão dois métodos considerados *baseline* imparcial, são eles o WKNNI e o EC, seguidos do KMI e do KNNI e por último está o MC, que faz a imputação por média/moda sem particionar o conjunto de dados.

5.4.2.2 RESULTADOS PARA O *WILSON'S NOISE RATIO*

A outra medida de desempenho analisada foi o *Wilson's Noise Ratio* calculada para cada base de dados, onde aplicou-se o teste de Friedman para avaliar as hipóteses. Reitera-se que o WNR foi normalizado, portanto quanto maior o valor, melhor o desempenho do método. A Tabela 7 apresenta os valores para o WNR normalizado obtidos pelos métodos de imputação em cada conjunto de dados, com os melhores resultados destacados em negrito e a classificação final obtida por meio do teste de Friedman.

Conforme análise dos resultados dispostos na Tabela 7, percebe-se que o método proposto obteve bons resultados, uma vez que alcançou os melhores valores do WNR em 14 das

Tabela 7 – *Wilson's noise ratio* normalizado e o ranqueamento obtido a partir do teste de Friedman.

Bases	**Métodos de tratamento de valores ausentes**						
	CMC	EC	KMI	KNNI	MC	WKNNI	GAImp
AUD	73,874	69,82	72,072	70,721	71,171	70,721	**77,928**
AUT	67,391	71,739	69,565	69,565	65,217	69,565	**73,913**
BAN	85,057	83,333	86,782	85,057	86,782	84,483	**90,23**
BRE	44,444	33,333	44,444	**55,556**	44,444	**55,556**	**55,556**
CLE	**83,333**	66,667	**83,333**	66,667	66,667	**83,333**	**83,333**
CRT	89,189	81,081	89,189	86,486	89,189	86,486	**94,595**
DER	**100**	**100**	**100**	**100**	**100**	**100**	**100**
HEP	88	89,333	84	84	86,667	81,333	**90,667**
HOC	**91,413**	82,271	87,258	85,596	85,596	85,596	89,474
HOV	95,567	94,089	95,074	96,059	93,596	96,059	**96,552**
LUN	**40**	20	20	20	20	20	**40**
MAM	91,603	83,206	86,26	83,206	85,496	84,733	**93,13**
POS	**66,667**	**66,667**	**66,667**	**66,667**	**66,667**	**66,667**	**66,667**
PRT	60,87	56,522	57,488	55,072	53,623	55,072	**61,353**
WIS	**93,75**	87,5	87,5	87,5	**93,75**	87,5	**93,75**
RANK	**3,09**	**5,334**	**3,87**	**4,73**	**4,63**	**4,6**	**1,73**

15 bases. Este fato evidenciou-se pela classificação no ranking obtido pelo teste de Friedman, pois o GAImp obteve 1.73, seguido pelo CMC (3.09) e KMI (3.87). Novamente, três conjuntos de dados figuram na lista de pouca ou nenhuma variabilidade, são eles *dermatology* e *post-operative*, que também pertenciam à lista anterior (acurácia do classificador) e *lung-cancer*. Como hipóteses de tal fato, destaca-se a combinação entre a pouca quantidade de instâncias com valores ausentes (tais bases têm apenas 8, 3 e 5, respectivamente) e a incidência de VA em atributos com pouca correlação com o rótulo.

A Tabela 8 também mostra os *p*-valores ajustados por dois testes *post-hoc*, o Holm e Shaffer, ambos para intervalo de confiança de 95% ($\alpha = 0,05$), sendo que o procedimento de Holm rejeita as hipóteses que tem p-valor ≤ 0.00294 e o procedimento de Shaffer rejeita as hipóteses em que o p-valor é ≤ 0.00238. O último procedimento - Bergmann - rejeita as seguintes hipóteses:

- EC vs. GAImp
- KNNI vs. GAImp
- MC vs. GAImp
- WKNNI vs. GAImp

Portanto, analisando os resultados dispostos na Tabela 8 e as hipóteses avaliadas pelos procedimentos citados, é possível afirmar que o GAImp é estatisticamente superior aos métodos de imputação EC, MC, WKNNI e KNN. Em relação ao CMC e o KMI, não encontrou-se evidência estatística da superioridade em relação ao WNR. Contudo, apenas o CMC figurou na

Tabela 8 – p-valores ajustados pelos procedimentos *post-hoc* Holm e Shaffer com $\alpha = 0,05$.

Métodos de TVA	p	Holm	Shaffer
EC vs, GAImp	5,0228E-6	0,00238	0,00238
KNNI vs, GAImp	1,4284E-4	0,0025	0,00335
MC vs, GAImp	2,3652E-4	0,002631	0,00335
WKNNI vs, GAImp	2,7888E-4	0,00278	0,00335
CMC vs, EC	0,00463	0,00294	0,00335
KMI vs, GAImp	0,00684	0,00313	0,00335
CMC vs, KNNI	0,03839	0,00335	0,00335
CMC vs, MC	0,05191	0,00357	0,00357
CMC vs, WKNNI	0,05722	0,00384	0,00384
EC vs, KMI	0,06298	0,00417	0,00417
CMC vs, GAImp	0,08317	0,00455	0,00455
KMI vs, KNNI	0,27189	0,005	0,005
CMC vs, KMI	0,33108	0,00556	0,00556
KMI vs, MC	0,33108	0,00625	0,00625
EC vs, WKNNI	0,35254	0,00714	0,00714
KMI vs, WKNNI	0,35254	0,00833	0,00833
EC vs, MC	0,37485	0,01	0,01
EC vs, KNNI	0,44687	0,0125	0,0125
KNNI vs, WKNNI	0,86577	0,01666	0,01666
KNNI vs, MC	0,89911	0,025	0,025
MC vs, WKNNI	0,96629	0,05	0,05

lista competitivos ao GAImp, sendo que este método possui as ressalvas de ser dependente de rótulo e de inserir um viés desnecessário às análises.

5.4.2.3 DISCUSSÕES

Apesar de já destacado nas subseções prévias, algumas análises merecem aprofundamento, primeiramente no tocante ao desempenho do GAImp. De acordo com os resultados para as medidas de desempenho analisadas (acurácia do classificador e *Wilson Noise Ratio*) e as evidências estatísticas obtidas, é possível afirmar que o método proposto obteve melhor desempenho que os demais métodos de imputação, oferecendo o melhor balanço na otimização das duas medidas adotadas, uma vez que sempre obteve boa posição nos *rankings* estatísticos - com a ressalva supramencionada ao CMC.

Em relação à parametrização, os valores dispostos na Tabela 4 apresentaram um balanço entre o tempo de processamento e a qualidade dos resultados. Os testes iniciais de calibração visaram primeiramente avaliar o impacto do tamanho da população e do número de gerações na qualidade da solução e na convergência. Como conclusões, destacam-se:

- Neste nicho de aplicação, a qualidade do resultado é mais sensível ao tamanho da população de forma diretamente proporcional; i.e. quanto maior a população, melhor o desempenho do GAImp em relação às acurácias dos classificadores e do WNR. Esta conclusão está em concordância com a demonstração da eficiência da imputação múltipla feita por Little e Rubin (1987) (Vide Eq.5.1);

- custo computacional está diretamente relacionado ao número de consultas a função de aptidão (construção dos modelos de classificação), seu número é determinado pelo tamanho da população multiplicado pelo número de gerações, portanto um equilíbrio entre estes dois parâmetros faz-se necessário;

- melhor *trade-off* entre convergência, custo computacional e qualidade da soluções apresentou-se com o número de gerações inferior ao tamanho da população.

Os demais parâmetros foram analisados com pouca variação devido seu baixo impacto no desempenho do algoritmo, por exemplo, para o número de indivíduos elitistas, dois valores foram testados, 1 e 3, sendo o último o selecionado para compor a configuração do algoritmo. Vale ressaltar que os valores escolhidos estão de acordo com a literatura da área.

Conforme explicitado anteriormente, o desempenho computacional está intimamente relacionado ao número de consultas à função de aptidão e ao algoritmo de classificação. Neste ponto, diversos algoritmos foram testados para compor o *framework* experimental. Dos adotados no estudo, o LWL demandou maior tempo de processamento, chegando a ser 10 vezes mais custoso que o Naïve Bayes. Devido ao caráter iterativo do método, alguns algoritmos aumentam demasiadamente o tempo de processamento sem trazer ganhos consideráveis à qualidade das soluções, com destaque ao SMO (aproximadamente 10 vezes mais custoso que o LWL).

5.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi estabelecido um paralelo entre a imputação múltipla de dados e algoritmos evolucionários para então apresentar o GAImp, um método de imputação de dados baseado em algoritmos genéticos. Após análise dos trabalhos correlatos, estabeleceu-se as seguintes metas para o método proposto: tratar conjuntos de dados com atributos de tipos mistos de forma satisfatória; considerar o advento de instâncias incompletas e informações da geração do modelo, mais especificamente a acurácia do classificador.

A análise de desempenho do GAImp foi realizada em um *framework* experimental composto de seis métodos, escolhidos a fim de representar três grupos de métodos de classificação: indução de regras, modelos de aproximação e aprendizado baseado em instâncias; os conjuntos de dados escolhidos já contém valores ausentes e apresentam atributos categóricos, numéricos e mistos. O GAImp foi comparado em relação a sete métodos de imputação, utilizando como medidas de desempenho a acurácia do classificador e o *Wilson Noise Ratio*.

As evidências estatísticas apontam que o método proposto obteve melhores resultados quando as métricas são analisadas em conjunto. Mesmo com a obtenção de resultados inferiores ao CMC, o GAImp utiliza-se de múltiplas imputações e de um processo iterativo para refiná-las, refletindo a variabilidade amostral sobre um modelo de não resposta ou incerteza acerca do modelo correto para a não resposta, em detrimento da imputação de um valor único, calculado

a partir dos dados de rótulo, tal como o CMC. Como restrição, chama-se atenção à iteratividade do método e da necessidade de construção de classificadores para cada indivíduo em todas as gerações, portanto, seu custo computacional torna o método inviável para grandes bases de dados. A descrição da proposta e resultados parciais foram publicados na forma de artigo em conferência internacional (LOBATO et al., 2015b); e a extensão das análises foram submetidas à um *special issue* sobre otimização combinatorial, contendo também alguns detalhes sobre a formalização do problema de imputação apresentado no Capítulo 3.

Destaca-se também que apesar do estudo de caso analisado ser a classificação de padrões, chama-se atenção para a flexibilidade da proposta, a qual pode ser adaptada a outras tarefas que envolvam análise multivariada, como regressão, classificação multirrótulo, agrupamento e análise de séries temporais; também possibilitando otimizar múltiplas medidas de desempenho, inclusive conflitantes, por meio da adoção de estratégias multiobjetivo, tal como será abordado no próximo capítulo.

6 MOGAIMP: ALGORITMO GENÉTICO MULTIOBJETIVO PARA IMPUTAÇÃO MÚLTIPLA DE DADOS

6.1 CONSIDERAÇÕES INICIAIS

Em razão das diversas medidas de desempenho para se avaliar um método de imputação e das características do GAImp, sobretudo sua flexibilidade e semelhança à imputação múltipla, vislumbrou-se a possibilidade de adaptá-lo para considerar múltiplas medidas de desempenho. Como algumas destas medidas mostram-se conflitantes, pois quando otimiza-se uma o desempenho da outra decai, percebeu-se a necessidade de adotar uma abordagem multiobjetivo, a qual ainda não havia sido explorada na literatura de tratamento de valores ausentes.

Neste capítulo, o algoritmo genético multiobjetivo para imputação de dados, denominado MOGAImp é apresentado. Como extensão do GAImp, este método também herda algumas propriedades como a consideração de informações provenientes da construção do modelo, utilização dos registros incompletos para estimar os valores a serem imputados, além de ser adequado para utilização em conjuntos de dados com atributos mistos.

As medidas a serem otimizadas pelo MOGAImp foram escolhidas pois são notoriamente conflitantes, a saber: manteve-se as acurácias dos classificadores, tal como na abordagem mono-objetivo proposta, e adicionou-se a acurácia preditiva dos métodos de imputação. Além do caráter conflitante, adotou-se esta segunda por possibilitar a emulação dos VA e posterior extração de um modelo de "regras de imputação" a partir dos valores preditos pelo MOGAImp.

Devido às características das medidas de desempenho para imputação de dados discutidos pelos trabalhos comparativos discutidas no Capítulo 4, e pelas abordagens multiobjetivo descritas no Capítulo 2, optou-se por usar a metodologia baseada em fronteiras de Pareto; pois por meio dela é possível melhor analisar o comportamento e a relação entre as medidas de desempenho, bem como selecionar soluções não dominantes. Mais especificamente, escolheu-se como técnica-base do MOGAImp um dos algoritmos para exploração dos conjuntos de Pareto mais utilizados no âmbito da mineração de dados, o NSGA-II. Esta escolha deu-se por este algoritmo ser elitista, eficiente computacionalmente e de fácil parametrização (DEB et al., 2002). Adicionalmente, com o intuito de reduzir o tempo de processamento, o MOGAImp também incorpora um esquema de paralelismo para o cálculo das funções de aptidão.

O *framework* experimental foi ligeiramente modificado para possibilitar a adição da acurácia preditiva do método de imputação, calculada a partir da distância entre o valor real da base e o valor imputado. Para tal, 15 conjuntos de dados com valores ausentes induzidos foram utilizados nos experimentos; cinco classificadores, representando os três grupos de métodos de classificação, foram selecionados para prover informações acerca da construção do modelo; e

comparou-se o desempenho do método proposto contra três métodos de imputação. Por fim, os resultados são apresentados e discutidos para posterior apresentação das considerações finais.

6.2 MÉTODO PROPOSTO: MOGAIMP

Tal como na abordagem mono-objetivo, os agentes motivadores para utilização de algoritmos genéticos no MOGAImp são: i) adaptabilidade à imputação múltipla; ii) a facilidade de codificação e paralelismo; iii) flexibilidade; e iv) auxílio na compreensão do problema. Aliado a estes fatores, tem-se também a utilização de medidas de desempenho conflitantes, justificando adoção de uma abordagem multiobjetivo. Devido às características do problema, com destaque para as medidas de desempenho a serem otimizadas; e das três abordagens multiobjetivo consideradas (atribuição de pesos aos objetivos, lexicográfica e fronteiras de Pareto), optou-se pela adoção da abordagem baseada na exploração de conjuntos de Pareto, na qual os algoritmos genéticos são consagrados, conforme visto no Capítulo 2.

Um dos algoritmos genéticos de exploração dos conjunto de Pareto mais relevantes à mineração de dados é o NSGA-II. Este algoritmo ordena as soluções e atribui uma distância de multidão entre os indivíduos. Ambos os procedimentos baseiam-se na dominância de suas soluções, o que também guia o processo de seleção em direção a um espalhamento uniforme das soluções em uma Fronteira de Pareto. Considerando as características do seu funcionamento, este algoritmo foi eleito para compor o método proposto, o qual é detalhado a seguir. Primeiramente é apresentada a representação do indivíduo e as funções de aptidão para então adentrar na descrição do algoritmo.

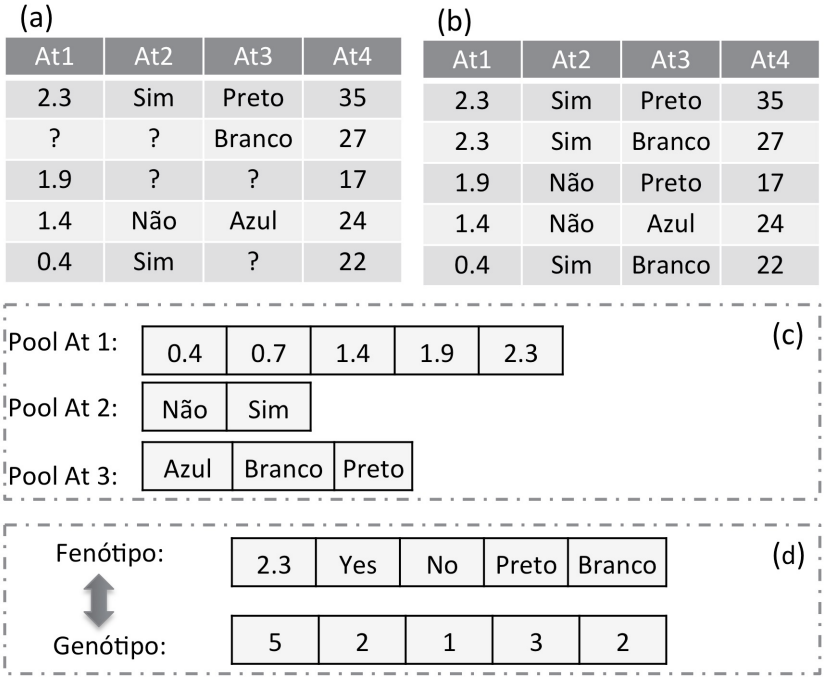
6.2.1 CODIFICAÇÃO DO INDIVÍDUO

Antes de adentrar na descrição do algoritmo, é necessário apresentar o esquema de codificação do indivíduo. Nesta versão, ainda utiliza-se o conceito de *pool* de soluções discutido no Capítulo anterior, mas com algumas alterações pois, ao contrário do GAImp, que trabalha com subconjuntos, escolheu-se tratar cada valor ausente individualmente. Portanto, no MOGAImp as soluções candidatas são representadas por um vetor no qual cada gene contém um valor que irá substituir um valor ausente, este raciocínio é apresentado na Figura 17.

A Figura 17 (a) mostra um conjunto de dados com cinco valores ausentes, os quais serão substituídos por valores plausíveis, gerando um conjunto de dados completo, tal como apresentado na Figura 17 (b). No MOGAImp o *pool* de soluções é construído a partir de todo o conjunto de dados, por exemplo, na Figura 17 (c) é possível ver três *pools* de soluções, um para cada atributo com dados ausentes, contendo todos os valores possíveis, de forma ordenada - atributos categóricos são ordenados lexicograficamente - para cada atributo.

O processo de codificação e decodificação dá-se da mesma forma do método mono-objetivo, o genótipo é um vetor de inteiros que referencia um valor no *pool* de soluções do seu

Figura 17 – Representação esquemática da codificação do indivíduo no MOGAImp.



Fonte: Elaborada pelo autor.

respectivo atributo; e o valor referenciado assume o respectivo alelo no fenótipo. Por exemplo, na Figura 17 (d) o genótipo é representado pelo vetor { 5, 2, 1, 3, 2 } e é mapeado consultando o *pool* de soluções do atributo o qual o alelo é correspondente. No caso do primeiro alelo, ele mapeará a partir do *pool* de soluções do At. 1, como resultado tem-se o mapeamento 5 (genótipo) = 2.3 (fenótipo).

Esta estratégia de codificação foi desenvolvida tendo em mente dois objetivos principais: i) prover uma abstração aos tipos de dados, permitindo a manipulação de atributos nominais e contínuos da mesma forma; ii) prover uma estrutura de dados adequada à aplicação de operadores genéticos e otimizar as funções de aptidão, as quais são apresentadas na subseção a seguir.

6.2.2 FUNÇÕES DE APTIDÃO

Como mencionado anteriormente, as medidas utilizadas para compor as funções objetivos foram pensadas de forma a possibilitar a extração de "regras de imputação" por meio da aplicação dos seguintes passos:

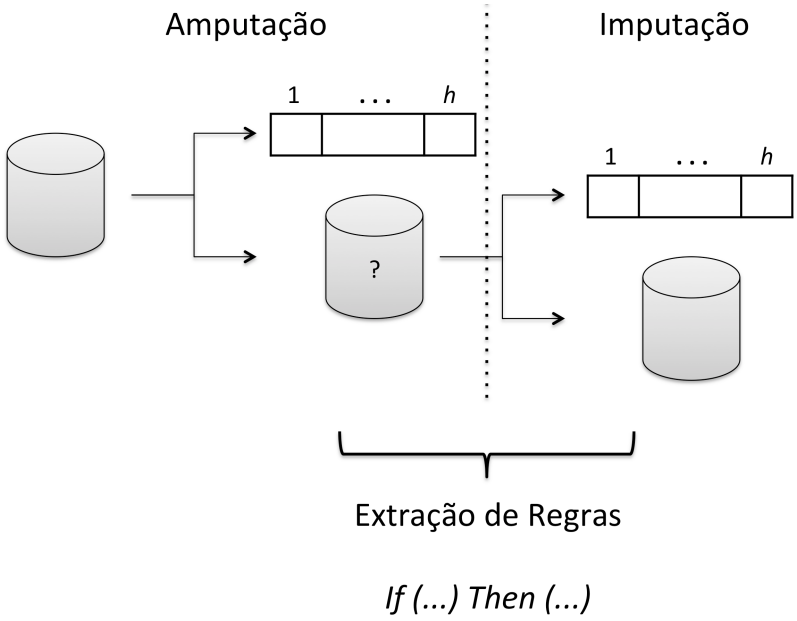
1. **Emulação da ausência de dados:** a maior parte dos experimentos envolvendo tratamento de valores ausentes considera que o mecanismo de ausência de dados que rege o conjunto de dados em questão é MCAR. Este passo consiste em remover valores observados de forma aleatória, a fim de induzir valores ausentes obedecendo o mecanismo de ausência completamente aleatório;

2. **Busca de valores ótimos para a imputação:** dadas as diversas medidas de desempenho, muitas das quais conflitantes, este passo busca por valores a serem imputados de forma a otimizar as referidas medidas;

3. **Extração de "regras de imputação":** a partir dos valores ótimos para as medidas de desempenho consideradas, extrair "regras de imputação" para serem utilizadas em tempo de classificação.

A Figura 18 representa graficamente estes passos. Na primeira parte tem-se a indução dos valores ausentes por meio da amputação - remoção de valores existentes visando induzir os valores ausentes - seguido da busca de valores que otimizem as medidas de desempenho escolhidas, é neste ponto que o MOGAImp é aplicado.

Figura 18 – Representação esquemática dos processos utilizados na análise do Grupo 2.



Fonte: Elaborada pelo autor.

Conforme visto na Figura 18, os processos de amputação e imputação geram dois produtos cada, um conjunto de dados e um vetor. O vetor advindo do processo de indução dos valores ausentes armazena os valores reais que foram removidos, enquanto o vetor proveniente do segundo processo contém os valores imputados. A partir destes vetores, calcula-se a distância entre o valor real e o valor imputado de acordo com a Eq. 6.1.

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(e_i - \overline{e}_i)^2} \tag{6.1}$$

Onde e_i é o i-ésimo valor original que foi removido da base e $\overline{e}_i$ e h é o número total de valores ausentes do conjunto de dados, sendo que para atributos categóricos considera-se o

resultado de $(e_i - \overline{e}_i)^2$ igual a 1. Para fins de legibilidade, adotou-se a normalização *min-max* por base, que retorna um valor entre 0-1, o qual invertido, resulta na medida (*Normalized Root Mean Square Error* (NRMSE)), que quanto maior, melhor a acurácia preditiva do método de imputação. O NRMSE foi escolhido para compor a função de aptidão do MOGAImp por ser fortemente presente em diversos trabalhos como forma de mensurar a acurácia preditiva, o que facilita a comparação posterior com métodos do estado da arte.

A segunda medida de desempenho adotada busca considerar informações obtidas a partir da construção do modelo de análise. Para o estudo de caso escolhido, a classificação de padrões, optou-se por manter a função de aptidão baseada na média das acurácias de l classificadores (Eq. 5.2), tal como no GAImp. Dessa forma nenhum método de imputação é beneficiado, uma vez que sabe-se a correlação entre desempenhos de método de imputação e algoritmos de classificação (LUENGO; GARCÍA; HERRERA, 2012; SIM; KWON; LEE, 2015).

Embora já discutido, faz-se importante pontuar algumas características acerca do RMSE: i) esta medida é sensível à quantidade de VA, pois durante o processo de amputação valores "únicos" podem ser perdidos no processo, portanto, ele não poderá ser inferido por nenhum método de imputação uma vez que não há registros com valor semelhante, consequentemente a diferença entre qualquer valor imputado e o valor real será alta; ii) o RMSE é considerado uma medida conflitante com a acurácia do classificador, em razão de que enquanto observa-se a otimização de uma, o desempenho da outra medida decai - a relação entre estas medidas é evidenciada na Seção 6.3.2.

6.2.3 FLUXO DE EXECUÇÃO

Descritas as estruturas de dados geradas, o esquema de codificação do indivíduo e as medidas de desempenho utilizadas como funções de aptidão, tem-se o subsídio necessário para apresentar o fluxo de trabalho do MOGAImp, destacando os operadores genéticos e o esquema de paralelismo adotado. O fluxo de trabalho baseia-se no NSGA-II e encontra-se descrito no Algoritmo 2, o qual tem como entradas um conjunto de dados incompleto e os parâmetros do algoritmo genético e retorna como resultado conjuntos de dados imputados. Definidas as entradas e saídas, monta-se a estrutura de dados necessária, o *pool* de soluções (Linhas 1-3).

O restante do algoritmo segue a estrutura tradicional do NSGA-II, começando pelo início da população, a qual é feita de forma aleatória e então os valores contidos nos indivíduos são imputados, cada um gerando um conjunto de dados completo (Linhas 4-5). A Linha 6 ilustra a avaliação da população de acordo com as funções objetivo adotadas, a média das acurácias dos classificadores e o RMSE. Usando as informações obtidas pelas etapas anteriores, a população inicial é submetida aos procedimentos de ordenamento da população e a atribuição da distância de multidão do NSGA-II, os quais organizam a população baseados na dominância das funções de aptidão e também guiam o processo de seleção rumo a um espalhamento uniforme das soluções em uma fronteira de Pareto ótima. No MOGAImp a seleção ocorre por torneio.

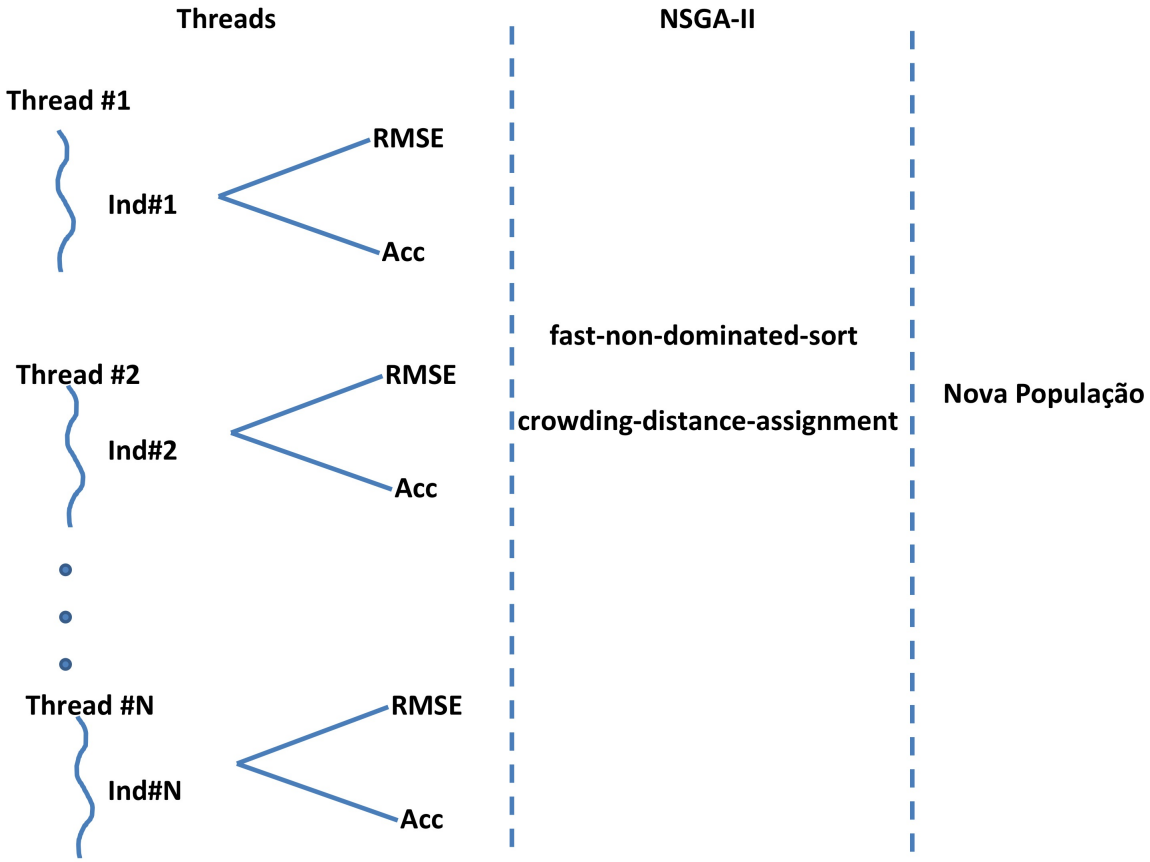
Algoritmo 2: Algoritmo genético multiobjetivo para imputação de dados

Entrada: Base de dados com VA, Parâmetros do AG
Saída: Bases imputadas
1 **para** *cada atributo k do conjunto de dados com VA* **faça**
2 | $pool(k) \leftarrow$ listar os valores do atributo (k);
3 **fim**
4 $P_0 \leftarrow$ Inicializar ($population\_size$, $pool$);
5 $D_0 \leftarrow$ Bases imputadas (P_0, $pool$);
6 $[O_{acc}, O_{RMSE}] \leftarrow$ Avaliar os Objetivos (D);
7 $P_0 \leftarrow$ Aplicar o procedimento non-dominated sort do NSGA-II (P_0, O_{acc}, O_{RMSE});
8 **para** $t \leftarrow 1$ *to max_generations* **faça**
9 | $Q_t \leftarrow$ Aplicar os operadores genéticos (P_{t-1}, O_{acc}, O_{RMSE});
10 | $D_t \leftarrow$ Imputar os conjuntos de dados (Q_t, $pool$);
11 | $[O'_{acc}, O'_{RMSE}] \leftarrow$Avaliar os objetivos ($D_t$);
12 | $R_t \leftarrow P_{t-1} \cup Q_t$;
13 | $P_t \leftarrow$ Selecionar os sobreviventes por meio do NSGA-II (R_t, O_{acc}, O_{RMSE});
14 **fim**
15 $F \leftarrow$ Fronteira de Pareto (P_t)
16 **retorna** Bases Imputados (F, $pool$);

Subsequentemente, a população inicial percorre o processo evolucionário (Linhas 8-14), até um determinado número de gerações (critério de parada). Pares de indivíduos são selecionados em torneio para a aplicação dos procedimentos de cruzamento e mutação a fim de se gerar os descendentes Q_t. Também visando um equilíbrio entre as características de exploração do espaço de busca (*exploration*) e a exploração de pontos ótimos (*exploitation*), escolheu-se como operador de cruzamento o *n-point crossover* e para realizar a mutação, adotou-se a mutação Gaussiana (*creep mutation*). As últimas duas linhas do Algoritmo 2 são responsáveis por extrair as soluções ótimas da FP de acordo com a distância desejada. Usualmente utiliza-se a distância da origem como ponto de equilíbrio entre as funções objetivo avaliadas, e por imputar os conjuntos de dados que serão retornados ao final da execução.

Em diversas aplicações práticas, os operadores do algoritmo genético são bastante simples e têm impacto limitado no tempo de processamento, no entanto, o cálculo da função de aptidão normalmente requer mais recursos computacionais. No MOGAImp, para cada indivíduo, constrói-se l modelos de classificação e calcula-se o RMSE em cada geração, é neste ponto que reside o gargalo da abordagem proposta. Visando reduzir o tempo de processamento sem interferir nas propriedades de busca do método, o MOGAImp é dotado de paralelismo para cálculo da função de aptidão, conforme disposto na Figura 19.

Figura 19 – Representação esquemática do paralelismo do MOGAImp.



A Figura 19 mostra como o paralelismo do MOGAImp é implementado. Para cada indivíduo uma *thread* é utilizada para calcular o RMSE e para construir os modelos de classificação, então estas informações compõem a função de aptidão usada pelos procedimentos do NSGA-II para geração dos descendentes.

6.3 EXPERIMENTOS E DISCUSSÕES

Os experimentos realizados para avaliar o desempenho do método proposto foram conduzidos de forma semelhante aos experimentos efetuados para o GAImp. Esta escolha deu-se devido à relevância e robustez das análises realizadas no referido estudo, conforme discutido no Capítulo 4. O recurso computacional é idêntico, composto de um servidor com 12 processadores de 2,1 Ghz cada, 16 Gigabytes de memória RAM, rodando a distribuição Linux CentOS 6.5 e Máquina Virtual Java versão 7, *update* 65. A fim de reduzir o gargalo de entrada e saída, os arquivos referentes aos conjuntos de dados foram lidos/escritos em uma partição virtual em memória RAM, usada como disco rígido. Nos referidos testes, buscou-se avaliar não somente a qualidade das soluções obtidas, mas também a influência dos parâmetros, convergência do método e custos computacionais, conforme é visto adiante.

6.3.1 *FRAMEWORK* EXPERIMENTAL

Os testes foram conduzidos usando cinco métodos de classificação: C4.5, Conjective e OneR representando o aprendizado por indução de regras; Naïve-Bayes, pertencente à categoria de modelos de aproximação; e do aprendizado baseado em instâncias escolheu-se o KNN. Estes classificadores foram escolhidos devido seu custo computacional reduzido e boa representatividade. A Tabela 9 apresenta os parâmetros padrões utilizados.

Tabela 9 – Parâmetros dos classificadores.

Classificadores	Parâmetros
C4.5	Poda = sim, confiança = 0.25, instâncias por folhas = 2
Conjective	Pesos mínimos de instâncias = 2
OneR	Número mínimo de objetos = 6
Naïve-Bayes	Sem parâmetro
K-NN	K = 3, função de distância = euclideana

Ao todo, 15 conjuntos de dados com valores ausentes induzidos foram utilizados nos experimentos e possuem a seguinte composição: 10 conjuntos de dados encontram-se publicamente disponíveis no repositório do KEEL (ALCALÁ et al., 2010); 5 encontram-se disponíveis no repositório da Universidade da Califórnia, Irvine (LICHMAN, 2013). Originalmente, as bases do grupo 2 não possuem valores ausentes. Com o intuito de permitir as análises, aplicou-se o processo de amputação de dados, gerando 20 conjuntos de dados e o desempenho foi computado a partir da média deles. As Tabelas 10 e 11 resumem os dois grupos de conjuntos de dados.

Tabela 10 – Conjuntos de dados obtidos do repositório KEEL (ALCALÁ et al., 2010).

Nome da base	*Acrônimo*	*At.*	*At. VA*	*Cat.*	*Cat. VA*	*Num.*	*Num. VA*	*Inst.*	*Inst. VA*	*Inst. VA(%)*	*VA(%)*
australian	AUS	15	14	7	6	8	8	690	487	70.58	8.39
ecoli	ECO	8	7	1	0	7	7	336	162	48.21	7.85
german	GER	21	20	14	13	7	7	1000	800	80	8.57
iris	IRI	5	4	1	0	4	4	150	49	32.67	7.2
magic	MAG	11	10	1	0	10	10	1902	1107	58.20	8.17
newthyroid	NEW	6	5	1	0	5	5	215	76	35.35	7.44
pima	PIM	9	8	1	0	8	8	768	390	50.78	7.98
shuttle	SHU	10	9	1	0	9	9	2175	1217	55.95	8.09
satimage	SAT	37	36	0	0	36	36	6435	5638	87.77	8.75
wine	WIN	14	13	1	0	13	13	178	125	70.22	8.34

Tabela 11 – Conjuntos de dados induzidos a partir de bases disponíveis no UCI (LICHMAN, 2013).

Nome da base	*Acrônimo*	*At.*	*At. VA*	*Cat.*	*Cat. VA*	*Num.*	*Num. VA*	*Inst.*	*Inst. VA*	*Inst. VA(%)*	*VA(%)*
Contraceptive	CTR	10	1	8	0	2	1	1473	471	31,44	3,54
Glass	GLS	10	1	1	0	9	1	214	50	23,22	2,72
Lymph	LYM	19	1	16	1	3	0	148	482	32,11	1,86
Tic-Tac-Toe	TTT	10	2	10	2	0	0	958	4283	44,22	5,89
Vertebral-Column	VTC	7	1	1	0	6	1	310	107	34,00	5,40

Para treinar os classificadores, utilizou-se o método validação cruzada *k-fold* com 10 subconjuntos. O desempenho do MOGAImp foi comparado contra métodos de *baseline* para lidar com valores ausentes, a saber: MC; CMC; e o WKNNI. Apenas este último necessitou de parametrização, a qual adotou-se o valor padrão disposto na literatura (k = 10).

Três medidas de desempenho foram utilizadas: as acurácias dos classificadores, o *Wilson's noise ratio* e o NRMSE. Para mensurar a significância estatística obtida pelas comparações entre os métodos de imputação, o teste de Wilcoxon pareado com intervalo de confiança de 90% foi aplicado para as acurácias dos classificadores e para o NRMSE; e o teste de Friedman, com intervalo de confiança de 90% e o procedimento *post-hoc* de Nemenyi para o WNR. Os motivos da adoção destes testes foram discutidos no Capítulo 5. Em resumo, o teste de Wilcoxon, ao passo de não ser paramétrico, é considerado seguro e robusto para comparações estatísticas pareadas (DERRAC et al., 2011); adicionalmente, as diferenças entre as acurácias obtidas são pequenas, reforçando a indicação deste teste estatístico, sobretudo pela sua simplicidade.

Para parametrização do método, diversos experimentos foram conduzidos visando analisar o impacto dos parâmetros quantitativos no desempenho, tanto em termos das medidas de desempenho adotadas (convergência e diversidade) quanto em relação ao custo computacional.

A primeira observação da parametrização do MOGAImp é relacionada com seu esquema de codificação, como neste método cada valor ausente é tratado individualmente, o espaço de busca foi aumentado, implicando na necessidade de se aumentar tanto o número de indivíduos quanto o numero de gerações - fato também relacionado com a escolha dos algoritmos de classificação utilizados para compor uma das funções objetivo. A segunda observação diz respeito à heterogeneidade das bases de dados obtidas do repositório KEEL, tanto em quantidade de valores ausentes quanto na sua dimensionalidade. Por conseguinte, os conjuntos de dados foram agrupados e variaram-se dois parâmetros quantitativos (tamanho da população e número de gerações) de acordo com a necessidade de cada grupo. Os demais foram mantidos idênticos, a saber: taxa de mutação = 50%, taxa de cruzamento = 100%; número de indiví-

duos por torneio = 10. A Tabela 12 apresenta o tamanho da população e o número de gerações adotado para cada grupo.

Tabela 12 – Parâmetros do MOGAImp.

Grupo	Bases	População	Gerações
G1	iris wine newthryoid ecoli	400	2000
G2	australian germam pima	350	1500
G3	shuttle magic	250	500
G4	satimage	100	100

Para as bases geradas artificialmente a partir de bases sem valores ausentes extraídas do repositório UCI, adotaram-se os mesmos valores para taxa de mutação (50%), taxa de cruzamento (100%) e número de indivíduos por torneio (10), mas devido as características das bases, utilizaram-se 250 indivíduos para a população e 1500 gerações. Devido ao caráter estocástico dos algoritmos genéticos, os resultados apresentados a seguir consistem na média de cinco execuções independentes.

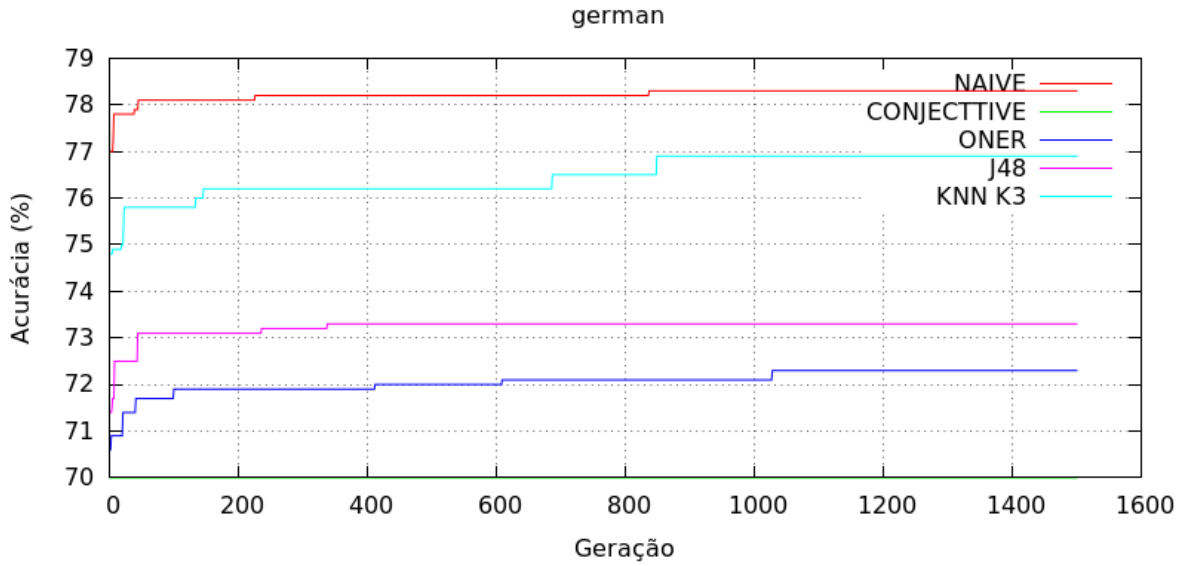
6.3.2 AVALIAÇÃO DE DESEMPENHO

Esta subseção apresenta os resultados obtidos considerando a metodologia experimental descrita, seguida de discussões acerca da relação entre as funções objetivo analisadas e a convergência do método. Três soluções foram extraídas da frente de Pareto, MOGAImp-RMSE, MOGAImp-ACC e MOGAImp-O, que representam respectivamente, as soluções com melhores RMSE, Acurácia e a com maior distância da origem, sendo um ponto de equilíbrio entre as duas funções objetivo consideradas.

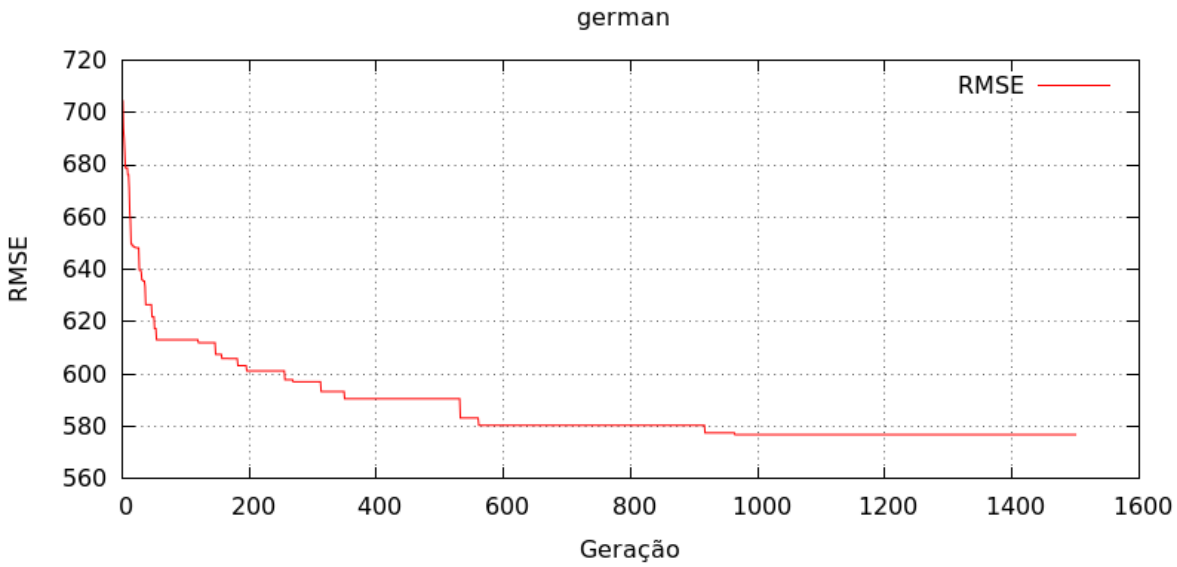
6.3.2.1 ANÁLISE PARA CONVERGÊNCIA

Um dos pontos pertinentes para avaliar o método proposto é analisar a convergência das soluções. Optou-se por avaliar as soluções mais distantes do ponto de origem (MOGAImp-O) por representar um equilíbrio entre as funções objetivo. As bases selecionadas para apresentação nesta seção foram *ecoli, german, magic e satimage*, pois são as mais representativas de seus grupos; sendo possível analisar o desempenho do melhor indivíduo no decorrer das gerações. As Figuras 20 - 23 apresentam as curvas de convergência em relação às acurácias dos classi-

ficadores (J48 representa o algoritmo C4.5) e ao RMSE para as bases *german, ecoli, magic e satimage*, respectivamente.



(a) Curva das acurácias.

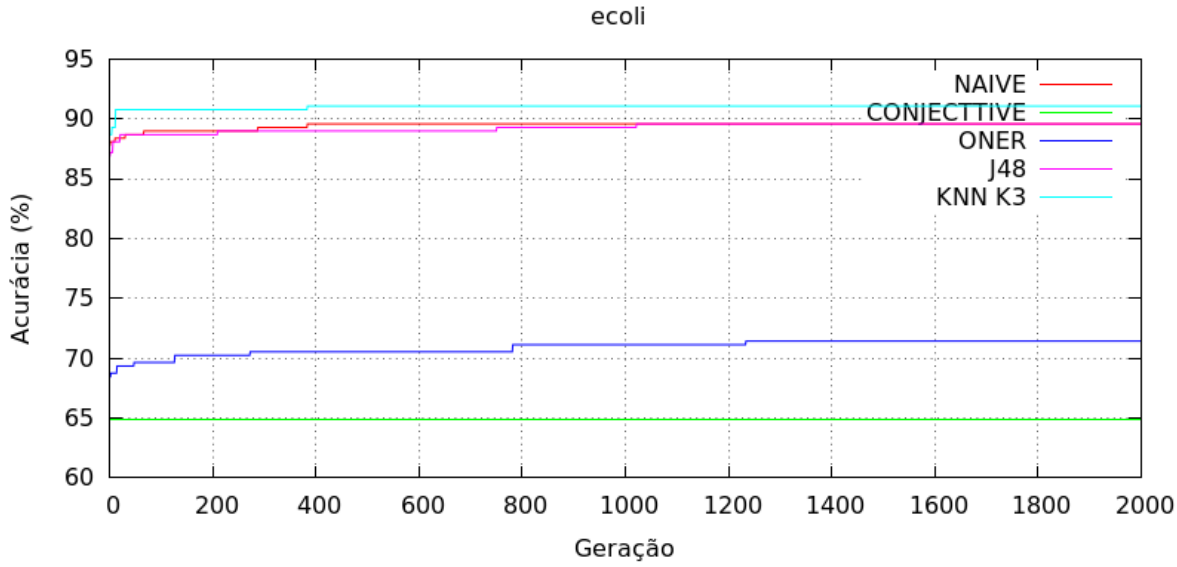


(b) Curva do RMSE.

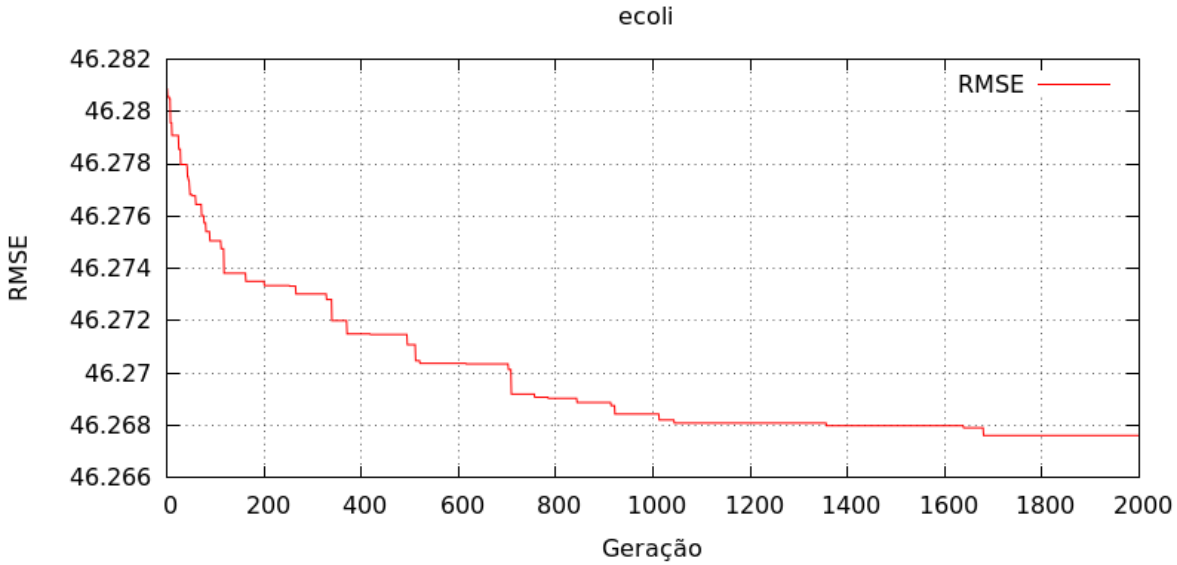
Figura 20 – Curvas de convergência para a base *german*.

Fonte: Elaborada pelo autor.

As Figuras 20a, 21a, 22a e 23a mostram a curva da acurácia para os conjuntos de dados em cada um dos classificadores adotados. É interessante verificar que o desempenho em relação aos classificadores varia consideravelmente nas bases analisadas, por exemplo, o classificador Conjective obteve boa acurácia para o conjunto de dados *german*, enquanto apresentou pior acurácia para a base *ecoli*, ratificando a necessidade de adoção de múltiplos classificadores a fim de assegurar a robustez do método.



(a) Curva das acurácias.



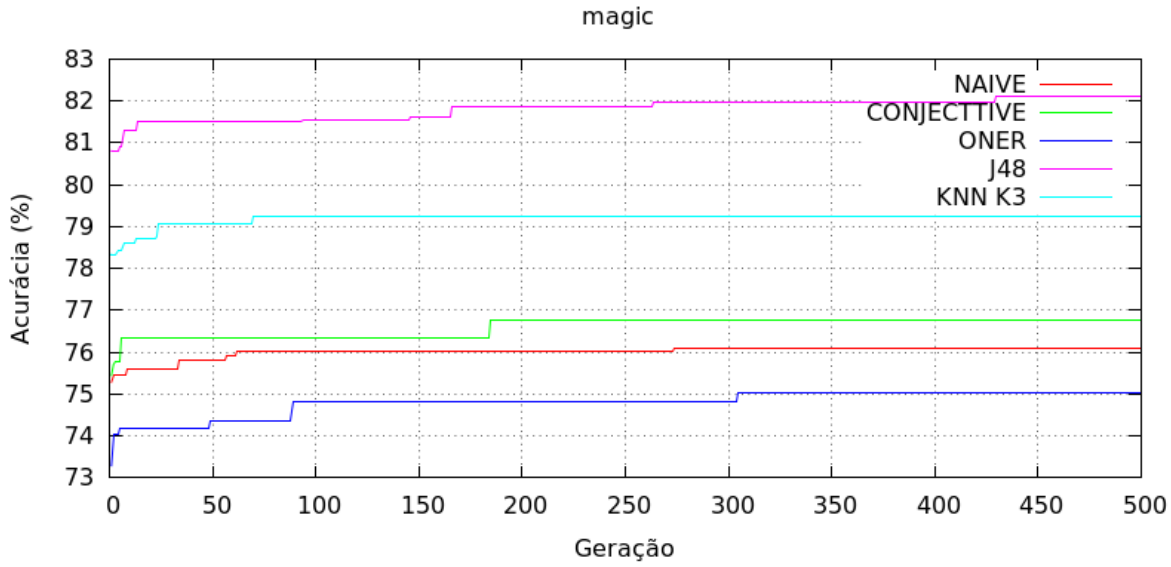
(b) Curva do RMSE.

Figura 21 – Curvas de convergência para a base *ecoli*.

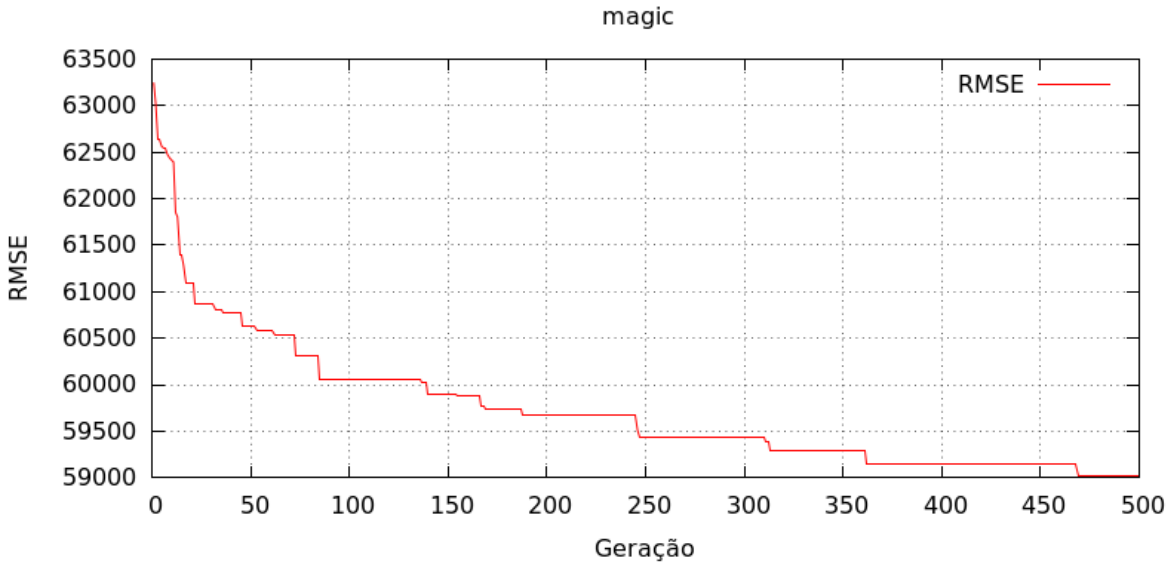
Fonte: Elaborada pelo autor.

Em relação ao RMSE, mostrado nas Figuras 20b, 21b, 22b e 23b, é possível observar mais mudanças em seu valor durante a execução e também diferentes escalas nas diferentes bases, uma vez que esta função objetivo está intimamente relacionada à quantidade de valores ausentes - quanto mais dados faltosos são presentes na base, maior o erro associado - e com os parâmetros do algoritmo. Para esta medida também são necessárias um maior número de gerações, no entanto ressalta-se o custo computacional envolvido com o aumento de parâmetros como tamanho da população e número de gerações.

Em comum às duas funções do objetivo, percebe-se que, até pelo seu caráter elitista,



(a) Curva das acurácias.

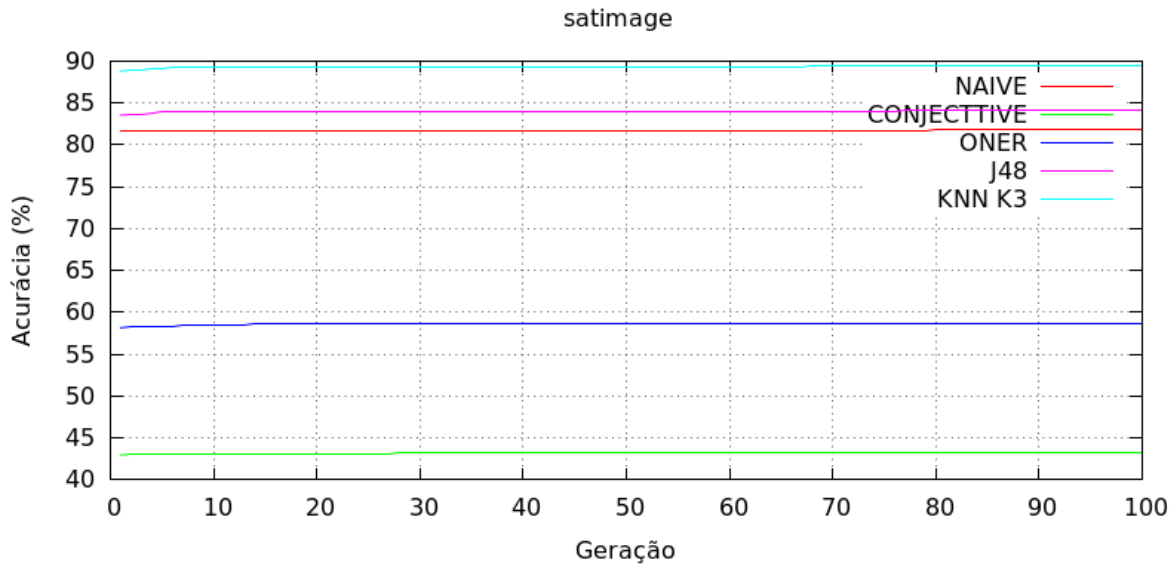


(b) Curva do RMSE.

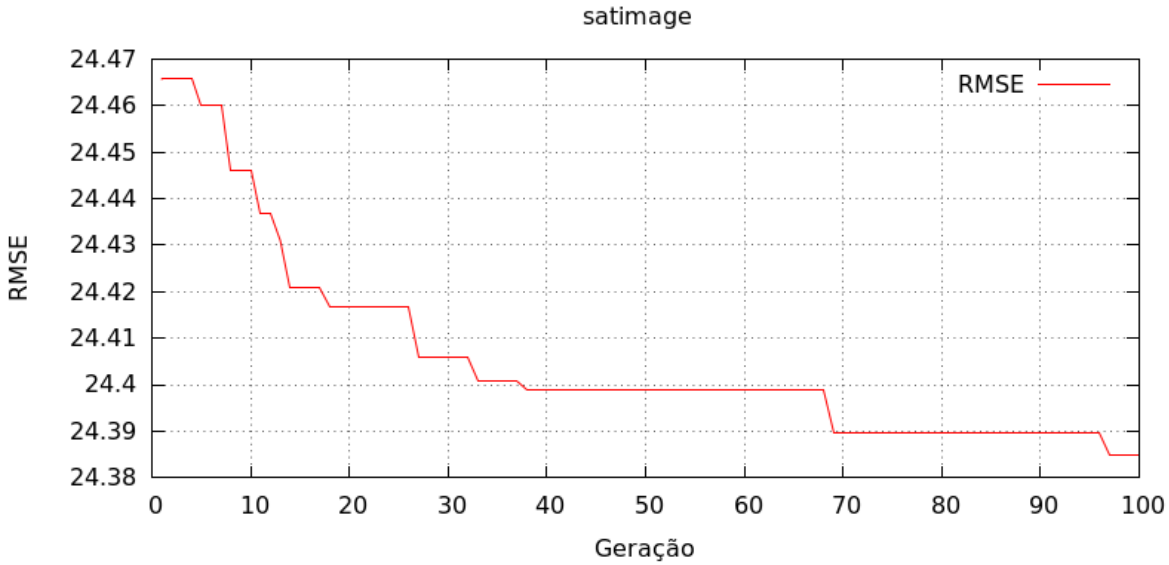
Figura 22 – Curvas de convergência para a base *magic*.

Fonte: Elaborada pelo autor.

ambas as medidas de desempenho convergem. No que tange à conflituosidade entre elas, as soluções MOGAImp-ACC e MOGAImp-RMSE evidenciam tal fato, como visto a seguir.



(a) Curva das acurácias.



(b) Curva do RMSE.

Figura 23 – Curvas de convergência para a base *satimage*.

Fonte: Elaborada pelo autor.

6.3.2.2 RESULTADOS PARA ACURÁCIA

A Tabela 13 apresenta o desempenho de cada método de imputação a respeito da acurácia dos classificadores utilizados levando em consideração os conjuntos de dados com valores ausentes induzidos. Os melhores resultados para a combinação entre algoritmo de classificação e os conjuntos de dados estão destacados em negrito.

Tabela 13 – Desempenho de cada método de imputação em relação à acurácia dos classificadores.

	Bases	**Métodos de Tratamento de Valores Ausentes**					
		MOGAImp RMSE	MOGAImp ACC	MOGAImp O	WKNNI	CMC	MC
CONJECTIVE	AUS	82,75	85,22	84,35	83,04	**86,67**	81,59
	CTR	43,08	43,44	43,38	43,04	**51,12**	42,98
	ECO	**64,88**	**64,88**	**64,88**	63,10	**64,88**	59,52
	GER	**70,00**	**70,00**	**70,00**	**70,00**	**70,00**	**70,00**
	GLS	44,29	**46,37**	45,22	44,39	45,64	44,39
	IRI	**66,67**	**66,67**	**66,67**	**66,67**	**66,67**	**66,00**
	LYM	74,02	76,20	74,62	72,45	**77,78**	72,60
	MAG	74,50	76,34	75,60	71,29	**76,60**	73,87
	NEW	78,60	**82,33**	80,93	74,42	77,67	77,21
	PIM	67,06	73,57	72,79	70,18	**73,70**	67,06
	SAT	42,74	43,20	43,17	43,62	**43,75**	41,80
	SHU	85,06	86,90	86,80	88,14	**88,55**	88,05
	TTT	67,91	69,61	68,52	67,35	**75,97**	68,26
	VTC	78,14	79,64	79,03	76,20	**80,00**	74,41
	WIN	64,61	**67,98**	66,85	63,48	63,48	60,11
C4.5	AUS	78,55	**86,38**	85,94	78,41	85,07	82,46
	CTR	47,76	53,13	51,29	47,92	**61,90**	47,27
	ECO	79,76	**88,99**	86,90	80,06	81,85	75,89
	GER	67,70	**73,40**	71,00	65,50	67,20	66,90
	GLS	69,31	**73,83**	71,86	67,24	69,42	67,03
	IRI	96,67	**98,67**	**98,67**	94,67	96,00	89,33
	LYM	77,55	**82,43**	80,56	75,90	79,13	77,10
	MAG	78,34	82,12	80,91	77,39	**83,07**	77,81
	NEW	92,09	**98,14**	97,21	91,63	92,56	90,23
	PIM	71,48	**78,78**	76,82	73,83	77,21	71,74
	SAT	82,32	83,99	83,99	85,50	**87,97**	83,05
	SHU	98,99	**99,63**	99,54	99,45	99,13	98,67
	TTT	78,60	84,45	82,56	78,23	**88,84**	78,32
	VTC	82,04	**88,21**	86,27	81,11	85,84	80,61
	WIN	90,45	**97,19**	96,07	89,89	91,57	89,89
KNN	AUS	80,29	**86,09**	84,78	79,57	85,51	80,87
	CTR	45,25	**49,14**	48,38	44,76	49,01	45,36
	ECO	86,01	**90,77**	88,99	81,85	86,61	82,14
	GER	71,60	**76,10**	74,60	72,00	71,60	72,10
	GLS	70,66	**74,04**	73,21	71,03	71,75	70,04
	IRI	96,00	**98,67**	**98,67**	95,33	96,00	92,00
	LYM	80,11	**85,59**	83,63	78,15	83,41	79,13
	MAG	75,39	**79,13**	78,55	76,39	78,13	76,13
	NEW	93,95	**97,67**	96,74	94,42	95,35	92,09
	PIM	72,01	**76,69**	74,22	73,05	72,66	70,05
	SAT	88,59	89,11	88,87	90,83	**91,11**	88,44
	SHU	98,02	**99,17**	98,80	99,45	98,99	98,44
	TTT	86,70	90,65	89,71	85,76	**94,20**	86,07
	VTC	80,68	**85,48**	83,87	78,78	82,54	77,53
	WIN	97,75	**99,44**	**99,44**	94,38	**99,44**	97,75

Tabela 13 – Continuação.

Bases		Métodos de Tratamento de Valores Ausentes					
		MOGAImp RMSE	MOGAImp ACC	MOGAImp O	WKNNI	CMC	MC
NAIVE-BAYES	AUS	74,78	78,70	77,68	76,38	**79,28**	75,94
	CTR	49,46	51,32	50,72	49,05	**52,32**	49,11
	ECO	84,82	**88,69**	87,50	80,95	85,71	83,04
	GER	74,50	**78,10**	76,70	74,50	75,40	75,10
	GLS	49,84	**54,62**	52,28	48,86	49,74	48,29
	IRI	96,00	**98,00**	**98,00**	95,33	**98,00**	87,33
	LYM	82,96	**85,36**	84,76	82,28	84,31	82,51
	MAG	75,03	**76,13**	75,66	72,29	73,24	72,87
	NEW	96,74	**98,14**	**98,14**	96,28	97,67	95,35
	PIM	76,95	**78,91**	77,99	75,13	76,69	74,22
	SAT	80,98	81,71	81,34	79,60	**82,28**	79,16
	SHU	92,74	**95,17**	93,84	93,15	93,43	90,21
	TTT	69,75	72,32	70,91	70,19	**78,87**	69,15
	VTC	79,43	**83,41**	81,86	76,85	80,36	76,56
	WIN	98,31	**99,44**	**99,44**	97,19	98,31	96,63
ONER	AUS	82,61	85,22	84,06	83,04	**86,67**	81,59
	CTR	46,50	48,85	47,39	45,90	**57,60**	46,64
	ECO	63,69	**70,24**	67,86	61,61	68,45	60,42
	GER	67,20	**72,10**	71,30	65,60	67,80	67,30
	GLS	59,09	**63,19**	60,49	57,11	60,12	54,78
	IRI	95,33	**97,33**	**97,33**	94,00	95,33	90,67
	LYM	74,32	75,38	75,15	74,02	**78,00**	73,50
	MAG	71,35	**74,76**	73,87	68,56	72,29	69,82
	NEW	92,09	**94,42**	93,95	91,16	91,16	88,84
	PIM	73,70	76,56	76,56	73,31	**77,21**	72,40
	SAT	57,89	58,60	58,55	60,14	**63,17**	56,43
	SHU	95,08	95,63	95,31	95,45	**95,77**	94,53
	TTT	68,05	69,59	68,56	67,32	**76,12**	68,59
	VTC	76,81	**82,62**	78,64	73,58	78,21	73,15
	WIN	82,02	**85,96**	84,83	75,84	79,21	72,47

Conforme esperado, dentre as outras soluções provenientes do método proposto, a solução MOGAImp ACC foi a que obteve melhor desempenho em relação a acurácia. Por meio da análise da Tabela 13, é possível perceber também que o método proposto encontrou dificuldade em otimizar a acurácia para o classificador Conjective, enquanto apresentou melhores resultados para os classificadores C4.5, KNN e Naïve Bayes. Também é possível observar que a solução MOGAImp-O figurou próximo aos melhores resultados encontrados.

Este fato é melhor visto por meio da distribuição e concentração das acurácias dispostas no boxplot apresentado na Figura 24, este gráfico ilustra de forma sutil a conflituosidade entre a acurácia do classificador e RMSE, conforme será melhor discutido na próxima seção.

Apesar de não ter obtido as melhores acurácias, por meio da análise da Figura 24 é possível perceber que as soluções obtidas pelo MOGAImp-O mostram-se competitivas em relação às demais. Fato ratificado na análise estatística dos resultados dispostos na Tabela 13, que contém os resultados da aplicação do teste pareado de Wilcoxon, onde obteve-se o ranqueamento

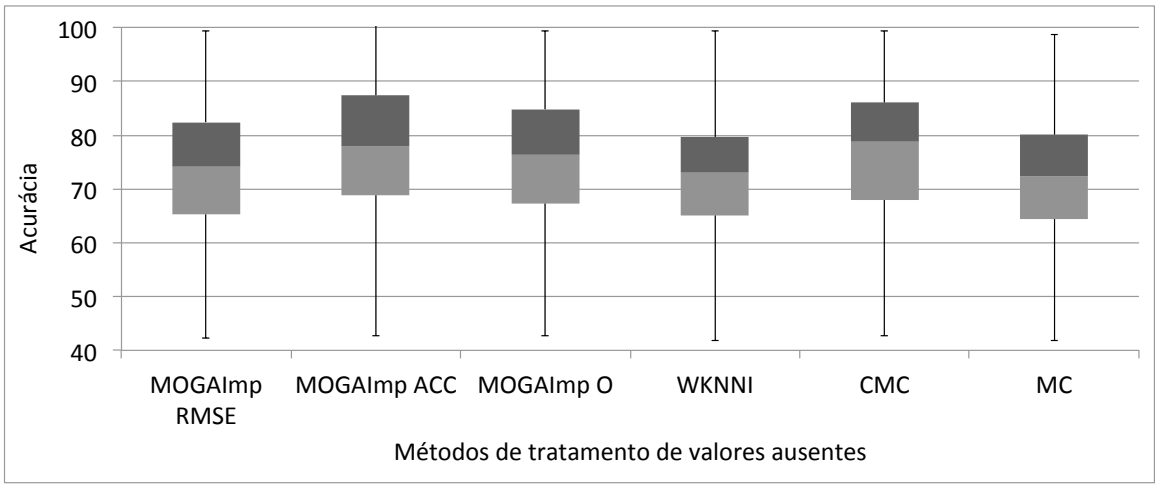


Figura 24 – Boxplot da acurácia dos classificadores nas comparações globais.

Fonte: Elaborada pelo autor.

dos métodos de imputação para cada classificador. A Tabela 14 apresenta os resultados do teste pareado de Wilcoxon.

Tabela 14 – Resultados do teste pareado de Wilcoxon para acurácia por método de classificação.

Imp. Methods	C4.5	KNN	Naïve-Bayes	Conjective	OneR	AVG	RANK
MOGAImp-RMSE	4	4	4	4	4	4	4
MOGAImp-ACC	1	1	1	2	1	1.2	1
MOGAImp-O	3	2.5	3	3	3	2.9	3
WKNNI	5.5	5.5	5.5	5	5	5.3	5
CMC	2	2.5	2	1	2	1.9	2
MC	5.5	5.5	5.5	6	6	5.7	6

Como esperado, o MOGAImp-ACC obteve melhores resultados para a acurácia dos classificadores, seguido do CMC e do MOGAImp-O. Em relação ao desempenho superior do CMC em relação ao MOGAImp-O, reitera-se que o CMC é dependente do rótulo, o que impede sua aplicação em tempo de classificação e que por ser um método de imputação simples, não reflete a variabilidade inerente da ausência de dados.

6.3.2.3 RESULTADOS PARA O RMSE

A Tabela 15 apresenta o resultado para o RMSE normalizado (calculado a partir da Eq. 6.1), representando a distância entre os valores reais e os imputados. Os melhores resultados estão marcados em negrito.

Por meio da análise da Tabela 15, é possível verificar o impacto da parametrização. As bases do Grupo 1 (iris, wine, newthryoid e ecoli), que apresentava população composta de 400 indivíduos evoluída por 2.000 gerações, foram as que obtiveram melhores resultados. Para as

Tabela 15 – Resultados do NRMSE por conjunto de dados.

Base	MOGAImp RMSE	MOGAImp ACC	MOGAImp O	WKNNI	CMC	MC
AUT	0,97	0,42	0,97	0,87	**0,98**	0,96
CTR	0,53	0,43	0,47	**0,72**	0,68	0,40
ECO	**0,98**	0,64	**0,98**	0,48	0,52	0,00
GER	**0,97**	0,19	**0,97**	0,29	0,95	0,65
GLS	0,79	0,50	0,62	**0,81**	0,69	0,49
IRI	**0,99**	0,78	0,93	0,88	0,92	0,00
LYM	0,35	0,20	0,23	0,45	**0,70**	0,49
MAG	0,84	0,15	0,84	0,56	**1,00**	0,92
NEW	**0,97**	0,30	0,88	0,49	0,65	0,66
PIM	0,92	0,21	0,92	0,59	**0,93**	0,53
SHU	0,84	0,14	0,84	0,88	**0,99**	1,00
TTT	0,11	0,05	0,08	0,54	**0,74**	0,53
VTC	0,71	0,52	0,62	**0,89**	0,71	0,62
WIN	**0,99**	0,58	0,97	0,04	0,76	0,00

demais bases do repositório do KEEL, onde executou-se o MOGAImp com valores inferiores para estes parâmetros, percebe-se uma queda de desempenho para o NRMSE.

Ainda acerca dos resultados do NRMSE dispostos na Tabela 15, não foi possível observar nenhuma relação entre conjuntos de dados com atributos com valores ausentes de tipo exclusivamente categórico ou numérico ou de tipos mistos. Em relação ao desempenho do WKNNI, este obteve bons resultados apenas para as bases induzidas. Em comum a estas bases, está o fato delas possuírem poucos atributos com VA, embora apresente em média 30% de instâncias com VA.

A Figura 25 mostra o boxplot do NRMSE para os métodos de imputação. Nesta figura é possível observar que a acurácia do classificador e o RMSE são medidas conflitantes, uma vez que as soluções que otimizam a acurácia (MOGAImp-ACC) apresentam queda no desempenho do RMSE. O mesmo ocorre quanto otimiza-se o RMSE (MOGAImp-RMSE), a acurácia dos classificadores não a acompanha (vide Figura 24).

Também aplicou-se o teste pareado de Wilcoxon para obter o ranqueamento dos métodos de imputação avaliando-se o NRMSE quando executou-se o MOGAImp por classificador, conforme apresentado na Tabela 16 - isto foi feito a fim de se verificar a influência dos algoritmos de classificação na otimização do RMSE.

Por meio da análise da Tabela 16, é possível perceber que o algoritmo de classificação não impacta na otimização do NRMSE, por exemplo, um determinado algoritmo de classificação beneficia a otimização do NRMSE.

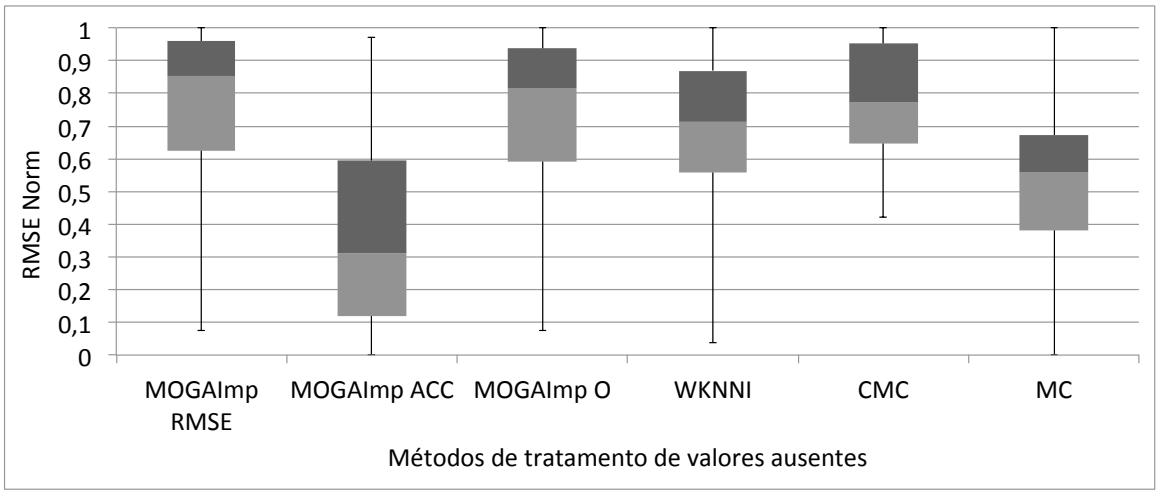


Figura 25 – Boxplot do NRMSE para os métodos de imputação de dados analisados.

Fonte: Elaborada pelo autor.

Tabela 16 – Resultados do teste pareado de Wilcoxon para acurácia por método de classificação.

Imp. Methods	C4.5	KNN	Naïve-Bayes	Conjective	OneR	AVG	RANK
MOGAImp-RMSE	1	1	1.5	1	1	1.1	1
MOGAImp-ACC	6	6	6	6	6	6	6
MOGAImp-O	4	4	4	4	4	4	4
WKNNI	2.5	2.5	3	2.5	2.5	2.6	3
CMC	2.5	2.5	1.5	2.5	2.5	2.3	2
MC	5	5	5	5	5	5	5

6.3.2.4 RESULTADOS PARA O *WILSON'S NOISE RATIO*

Os resultados para o *Wilson's noise ratio* normalizado, bem como o ranqueamento obtido pelo teste de Friedman, são apresentados na Tabela 17, com os melhores resultados destacados em negrito.

Por meio da análise dos resultados dispostos na Tabela 17, é possível perceber que, apesar de não obter os valores máximos, as soluções MOGAImp-ACC e a MOGAImp-O mostram-se competitivas no ranqueamento obtido pelo teste de Friedman. O CMC e o MOGAImp-ACC foram os que apresentaram melhor desempenho em relação ao WNR, pois esta métrica está correlacionada com a acurácia do classificador, sobretudo dos métodos de aprendizado baseado em instância.

Em relação ao desempenho estatístico, a solução MOGAImp-ACC e o método CMC são equivalentes e estatisticamente significantes, uma vez que estes métodos são estatisticamente superiores aos demais, conforme evidenciado na Tabela 18.

A Tabela 18 apresenta os *p*-valores ajustados por dois testes *post-hoc*, o Holm e Schaffer, ambos para o intervalo de confiança de 95% ($\alpha = 0.05$), sendo que o procedimento de Holm

Tabela 17 – *Wilson's noise ratio* normalizado e o ranqueamento obtido a partir do teste de Friedman.

Basest	**Métodos de tratamento de valores ausentes**					
	MOGA IMP RMSE	MOGA IMP ACC	MOGA IMP O	WKNNI	CMC	MC
AUS	85,05	87,02	85,05	85,01	**88,91**	85,01
CTC	61,86	62,75	62,77	60,66	**71,11**	61,16
ECO	88,89	91,23	90,49	82,72	**92,59**	84,57
GER	81,88	82,18	81,88	**82,88**	81,63	82,25
GLI	79,42	80,80	80,05	77,81	**83,56**	78,25
IRS	95,92	**97,14**	96,73	93,88	95,92	87,76
LPG	86,53	87,84	87,29	84,32	**88,42**	83,21
MAG	81,90	82,42	81,90	81,48	**85,19**	81,12
NTD	96,84	97,37	98,16	96,05	**98,68**	92,11
PIM	81,95	82,87	81,95	80,51	**84,87**	80,51
SAT	91,82	92,13	91,99	**93,56**	93,47	91,86
SHT	98,44	98,39	98,44	**99,43**	99,26	98,36
TTT	92,26	92,69	92,63	92,39	**94,07**	90,71
VTC	87,22	91,11	89,44	83,72	**96,23**	85,28
WNE	98,88	99,04	99,68	96	**100**	99,2
Ranking	**4,1**	**2,53**	**3,0**	**4,53**	**1,67**	**5,14**

$\{CMC, MOGAImp-ACC\} \succ \{MOGAImp-RMSE, WKNNI, MC\}$

Tabela 18 – *p*-valores ajustados pelos procedimentos *post-hoc* Holm e Shaffer para intervalo de confiança de 90%.

Métodos de TVA	*p*	Holm	Shaffer
CMC vs, MC	3,881478765029268E-7	0,0067	0,0067
WKNNI vs, CMC	2,712266493327232E-5	0,0071	0,01
MOGAImp-ACC vs, MC	1,4122651246579213E-4	0,0077	0,01
MOGAImp-RMSE vs, CMC	3,679909369104159E-4	0,0083	0,01
MOGAImp-O vs, MC	0,0021114910066706385	0,0091	0,01
MOGAImp-ACC vs, WKNNI	0,0034147911781178394	0,01	0,01
MOGAImp-RMSE vs, MOGAImp-ACC	0,021826990038418072	0,0112	0,0143
MOGAImp-O vs, WKNNI	0,028108040147151802	0,0125	0,0143
MOGAImp-O vs, CMC	0,045436036734246385	0,0143	0,0143
MOGAImp-RMSE vs, MOGAImp-O	0,11841994270812453	0,0167	0,0167
MOGAImp-RMSE vs, MC	0,13036982841080608	0,02	0,02
MOGAImp-ACC vs, CMC	0,2045587527205526	0,025	0,025
WKNNI vs, MC	0,3797754748409493	0,0334	0,0334
MOGAImp-ACC vs, MOGAImp-O	0,46421431277103115	0,05	0,05
MOGAImp-RMSE vs, WKNNI	0,5258621886847651	0,1	0,1

rejeita as hipóteses que tem p-valor ≤ 0.01112 e o procedimento de Shaffer rejeita as hipóteses em que o p-valor é ≤ 0.006667. O último procedimento, o de Bergmann rejeita as seguintes hipóteses:

- MOGAImp-RMSE vs. MOGAImpACC
- MOGAImp-RMSE vs. CMC
- MOGAImp-ACC vs. WKNNI
- MOGAImp-ACC vs. MC
- MOGAImp-O vs. MC
- WKNNI vs. CMC
- CMC vs. MC

6.3.2.5 DISCUSSÕES

Por usar um esquema de codificação onde cada valor ausente é tratado individualmente, tal como em outros métodos evolucionários para imputação de dados presentes na literatura, o MOGAImp requer uma parametrização diferenciada, sendo que dois parâmetros são os mais sensíveis para este domínio de aplicação: o número de indivíduos da população e o número de gerações, conforme observado nas análises referentes à convergência do método.

No entanto, o custo computacional associado ao aumento destes parâmetros deve ser levado em consideração, principalmente quando se pretende aplicar tais métodos em conjuntos de dados com complexidade considerável; como o caso da base *satimage*, a qual possui 37 atributos, todos eles apresentando valores ausentes, e mais de 6 mil instâncias das quais 87% estavam incompletas. Devido a estas características da *satimage*, o espaço de busca é grande, entretanto, o custo computacional para construção dos modelos de classificação (dimensionalidade) inviabilizam o uso de parâmetros adequados para o MOGAImp.

Esta correlação entre as características do conjunto de dados (dimensionalidade, distribuição dos valores ausentes *etc*), custo computacional para construção dos modelos (*e.g.* escolha de algoritmos de classificação) e a parametrização do método proposto abre possibilidades para: i) investigar estratégias para redução do espaço de busca sem utilizar agrupamento de instâncias; e ii) aplicar métodos para controle ou sintonia de parâmetros, automatizando esta tarefa.

Em relação ao desempenho do método proposto frente às medidas de desempenho adotadas, o MOGAImp mostrou-se competitivo e com potencial para aplicações reais devido sua flexibilidade, como por exemplo, incorporar múltiplas medidas de desempenho específicas de determinado nicho de aplicação; ou ainda, incorporar conhecimento de fundo por meio de restrições do tipo *cannot-link* ou *must-link*.

6.4 CONSIDERAÇÕES FINAIS

Neste capítulo, o algoritmo genético multiobjetivo para imputação de dados, denominado MOGAImp, foi apresentado como um método de imputação capaz de lidar com medidas de desempenho conflitantes. Este método representa uma extensão do GAImp, portanto herda algumas propriedades como a consideração de informações da construção do modelo, utilização dos registros incompletos para estimar os valores a serem imputados, além de ser adequado para utilização em conjuntos de dados com atributos mistos.

A análise de desempenho do método proposto levou em consideração três medidas de desempenho, o *Wilson's noise ratio* e outras duas que compunham as funções de aptidão do método proposto, a saber: as acurácias dos classificadores e a acurácia preditiva do método de imputação, calculada a partir da distância entre os valores reais e os valores imputados - estas medidas provaram-se conflitantes. Os resultados obtidos mostraram que o MOGAImp é competitivo, sua flexibilidade também merece ser destacada pois o método pode ser facilmente adaptado a outras tarefas de análise de dados (*e.g.* classificação multirrótulo, análise de séries temporais), por meio de pequenas modificações das funções objetivo; bem como a incorporação de conhecimento de fundo por meio da inclusão de restrições.

O método e as análises apresentadas neste capítulo foram publicadas em periódico da área de reconhecimento de padrões (LOBATO et al., 2015a), embora seu esquema de codificação e múltiplos objetivos requeiram valores para parâmetros quantitativos superiores aos do método mono-objetivo e consequentemente, maior custo computacional. Neste ponto, dois itens podem ser avaliados para reduzir o custo computacional por meio da diminuição do espaço de busca - isto graças à flexibilidade do método proposto - são eles: i) investigar a adoção de uma codificação de indivíduo baseada em agrupamento de instâncias ao invés de tratar cada valor ausente individualmente; e ii) utilizar métodos de imputação simples como soluções iniciais em detrimento da inicialização aleatória. Este último é um dos pontos abordados no próximo capítulo.

7 EXTRAPOLAÇÕES DOS MÉTODOS PROPOSTOS E ANÁLISES REALIZADAS

7.1 CONSIDERAÇÕES INICIAIS

Conforme evidenciado nos Capítulos 4, 5 e 6, a imputação de dados por meio de algoritmos evolucionários herda as características da imputação múltipla iterativa, o que a torna uma solução atrativa para diversos cenários. Por conseguinte, aliado à flexibilidade dos métodos apresentados nos Capítulos anteriores, é possível adaptá-los a diferentes cenários por meio de pequenas modificações, ou em sua codificação ou função de aptidão, extrapolando-os para outros problemas reais além da classificação de padrões tradicionais.

Neste ponto, dois cenários foram escolhidos: a análise de séries temporais, uma vez que esta tarefa de análise de dados comumente apresenta uma forte incidência de dados ausentes (HONAKER; KING; KING, 2013); e a classificação multirrótulo, pelo crescente interesse da comunidade de aprendizado de máquina neste tópico (ALVARES-CHERMAN; METZ; MONARD, 2012). Sendo assim, dois métodos de imputação específicos para estes domínios de aplicação são propostos.

O primeiro método, baseado em algoritmos genéticos e chamado MultImp, é aplicado à classificação multirrótulo, pois percebeu-se uma lacuna na literatura quanto a trabalhos que avaliassem o impacto de valores ausentes neste tipo de cenário. Em resumo, o MultImp representa uma extrapolação do MOGAImp, com as seguintes diferenças:

- **Abordagem multiobjetivo:** dado que medidas de desempenhos adotadas para avaliar a classificação multirrótulo não apresentam comportamento conflitante. Adicionalmente, visando diminuir o custo computacional, adotou-se no MultImp uma abordagem multiobjetivo lexicográfica;
- **Parametrização:** um dos gargalos identificados no MOGAImp é referente ao seu sistema de codificação, pois cada valor ausente é tratado individualmente, impactando diretamente no aumento do espaço de busca. Também com o objetivo de reduzir o custo computacional do método, adotou-se no MultImp uma estratégia para diminuição do espaço de busca por meio da inicialização de seus indivíduos utilizando soluções advindas de métodos de imputação simples.

O segundo é baseado no uso na programação genética, como método de regressão, para predizer os valores ausentes em cada atributo, aqui referenciado como GPImp. A habilidade desta técnica de computação evolucionária em aprender funções a partir de dados de exemplo fazem-na uma candidata em potencial para imputar dados em séries temporais, uma vez que a

maior parte dos dados deste tipo de análise é composta por atributos de tipo numérico. Outro item importante de salientar é que o método fornece modelos interpretáveis, uma vez que as funções de regressão podem ser facilmente visualizadas e interpretadas pelos especialistas do domínio. Estes métodos foram idealizados como extrapolações das análises realizadas e métodos previamente apresentados, pois:

- **Tarefa de análise:** os experimentos conduzidos até então haviam se concentrado na classificação de padrões. Com o intuito de extrapolar para outras tarefas, escolheu-se a análise de séries temporais, dado a relevância do tratamento de valores ausentes neste domínio de aplicação;
- **Interpretabilidade:** uma das limitações dos demais métodos propostos está relacionada com a interpretabilidade das soluções. Com o intuito te suplantar esta falha, optou-se pela extrapolação das análises visando a extração de regras de imputação, representada por funções de regressão.

Neste capítulo o MultImp e o GPImp e sua variante são descritos e contextualizados.

7.2 MÉTODO DE IMPUTAÇÃO MULTIOBJETIVO PARA OTIMIZAÇÃO DA CLASSIFICAÇÃO MULTIRRÓTULO

A classificação multirrótulo é um problema de aprendizado supervisionado onde uma instância pode estar associada a múltiplos rótulos, diferente da classificação tradicional que associa um exemplo a uma única classe (READ et al., 2011). O aprendizado multirrótulo é um tópico de pesquisa emergente e promissor devido ao número crescente de novas aplicações, como classificação semântica de vídeos e imagens, e categorização de música e texto (ALVARES-CHERMAN; METZ; MONARD, 2012); a exemplo, uma música pode ser categorizada como "*Blues*" e "Bossa nova" e um filme pode classificado como "Aventura" e "Animação".

A relevância da classificação multirrótulo motivou o desenvolvimento de um método de imputação para este nicho de aplicação, denominado de MultImp. Em decorrência da utilização de múltiplas medidas para avaliar o desempenho do aprendizado multirrótulo, o método proposto implementa uma abordagem multiobjetiva baseada em lexicografia, considerando três medidas bem estabelecidas: o casamento exato (*Exact Match* - EM), a acurácia e o *Hamming Loss* (HL) (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010). O cálculo destas depende de algumas definições, sendo que aqui adotaremos a notação apresentada por Goncalves, Plastino e Freitas (2013), a saber: n é o número de instâncias do conjunto de teste; q é o número de rótulos; Y_i é o conjunto de rótulos original da instância i; e Z_i é o conjunto de rótulos predito para a instância i.

De posse dessas informações é possível especificar as medidas de desempenho supracitadas. A *Exact Match* denota a taxa de predições em que todos os rótulos são previstos corretamente e é calculada conforme a Eq. 7.1.

$$EM = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \equiv Z_i) \tag{7.1}$$

Diferentemente da EM, a acurácia leva em considerações exemplos parcialmente corretos, em outras palavras, quando apenas um subconjunto dos rótulos do exemplo são corretamente preditos, por isso, é considerada uma medida mais flexível. A acurácia é calculada de acordo com a Eq. 7.2.

$$ACC = \frac{1}{n}\sum_{i=1}^{n} \frac{|Yi \cap Zi|}{|Yi \cup Zi|} \tag{7.2}$$

A última medida considerada é a *Hamming Loss*, que apresenta a média percentual das predições incorretas em relação ao número de rótulos, e pode ser calculada de acordo com a Eq. 7.3.

$$HL = \frac{1}{n}\sum_{i=1}^{n} \frac{Yi \Delta Zi}{q} \tag{7.3}$$

Para implementar a busca visando otimizar múltiplos objetivos, optou-se por uma abordagem baseada em lexicografia pelo seu custo computacional reduzido, em comparação com a exploração da fronteira de Pareto; e por evitar a especificação de pesos numéricos aos atributos. A ordem lexicográfica adotada foi: *Exact Match*, Acurácia e *Hamming Loss*. Apesar do MultImp ser baseado em algoritmo genético, o método proposto utiliza uma estratégia diferenciada para inicialização e controle da população, conforme abordado a seguir.

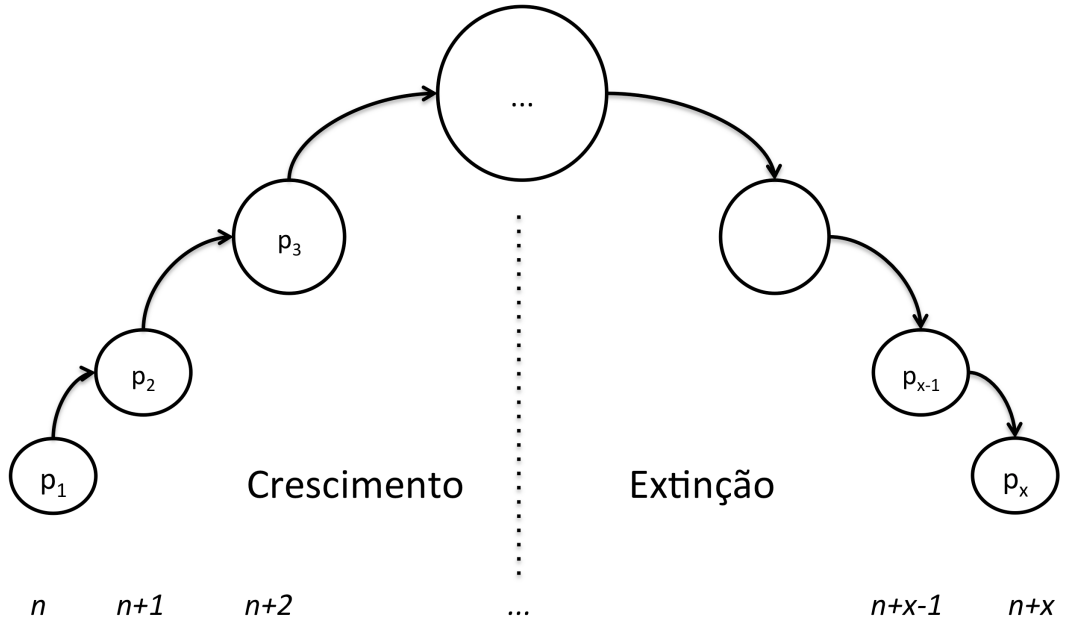
7.2.1 FUNCIONAMENTO DO MULTIMP

No MultImp, cada indivíduo gera um conjunto de dados completo (imputado), sua codificação é similar ao MOGAImp, onde cada valor ausente da base é um alelo. A fim de reduzir a desvantagem desse esquema de codificação, no tocante ao aumento do espaço de busca e consequente aumento no custo computacional, o método proposto implementa uma estratégia para inicialização, na qual os indivíduos da geração zero são obtidos a partir de métodos de imputação simples; para então aplicar o processo evolucionário, conferindo as propriedades de imputação múltipla iterativa ao MultImp.

Neste esquema de inicialização, utilizam-se j métodos de imputação simples como o KNNI e MC para gerar j indivíduos. A fim de compensar um possível tamanho de população reduzido, utiliza-se também uma estratégia determinística para controle deste parâmetro em tempo de execução, composta de duas fases: o aumento e a extinção. Na fase de aumento,

as gerações subsequentes são de tamanho superior à anterior, enquanto na fase de extinção, o tamanho da população é gradualmente reduzido. Esta lógica é apresentada na Figura 26.

Figura 26 – Representação esquemática do controle do tamanho da população em tempo de execução.



Fonte: Elaborada pelo autor.

Os círculos da Figura 26 denotam o tamanho da população, na fase de crescimento. O número de indivíduos vai aumentando até chegar no máximo, logo $p_1 < p_2 < .... < p_{\frac{x}{2}}$. Quando alcança-se a metade do número de gerações, passa-se para a fase de extinção onde o tamanho da população decresce na mesma taxa de crescimento da fase anterior. Durante esse processo, devido aos operadores genéticos selecionados, as soluções vão sendo combinadas até obterem um indivíduo que otimize as soluções de imputação simples, gerando uma solução mais robusta que seus geradores - princípio da imputação múltipla. Para tal, os seguintes operadores genéticos foram escolhidos: a seleção é realizada por torneio; o cruzamento é o *n-point-crossover* por proporcionar maior troca de material genético entre os indivíduos; e a mutação é baseada na substituição do valor do alelo por outro de um indivíduo selecionado aleatoriamente. Sendo assim, o MultImp é semelhante à imputação múltipla tradicional, com duas particularidades, sua inicialização a partir de métodos de imputação simples, e pela estratégia evolucionária empregada para combinar as soluções.

7.2.2 EXPERIMENTOS COMPUTACIONAIS

Os experimentos foram conduzidos utilizando quatro conjuntos de dados, os quais foram escolhidos por possuírem poucos rótulos, consequentemente, menor custo computacional para construir o modelo de classificação; e por não possuírem valores ausentes originalmente.

De forma a possibilitar as análises, induziu-se 5% de dados ausentes, obedecendo o mecanismo MCAR. A Tabela 19 apresenta um resumo dos conjuntos de dados utilizados.

Tabela 19 – Conjuntos de dados utilizados nos experimentos do MultImp.

Dataset	Domínio	(%VA)	Nº Inst.	Nº Atrib.	Nº Rótulos	(%) VA/Inst.	Total de VA
birds	audio	5	645	279	9	100	8997
emotions	música	5	593	78	72	76,29	2312
flags	imagens	5	194	26	7	100	252
CAL500	música	5	502	1213	174	100	6074

Para construção do modelo de classificação multirrótulo, escolheu-se o método *Binary Relevance* (BR), por ser bem estabelecido na literatura e por encontrar-se implementado no MULAN (TSOUMAKAS et al., 2011), biblioteca para classificação multirrótulo escrita em java. Como parâmetro do BR, informou-se o algoritmo de classificação utilizado para construir o modelo de classificação de cada rótulo, no caso, escolheu-se o C4.5, tal como em Goncalves, Plastino e Freitas (2013) e Gonçalves, Plastino e Freitas (2015).

Para comparação, utilizou-se dois métodos de imputação de dados, o KNNI com 10 vizinhos e o MC; utilizou-se também a acurácia sem imputação como *baseline*, uma vez que o C4.5 é um dos poucos algoritmos que implementa internamente uma estratégia para lidar com valores ausentes. Três medidas de desempenho foram adotadas: acurácia, EM e HL.

Os seguintes parâmetros foram utilizados para o MultImp: cinco indivíduos iniciais representando as soluções obtidas por KNNI, MC, EC, KMI e imputação aleatória; e sete gerações, com a seguinte lógica de crescimento/decrescimento da população sendo 5, 10, 25, 50, 25, 10 e 5, ou seja, cinco indivíduos na primeira geração, 10 indivíduos na segunda e assim sucessivamente.

7.2.2.1 RESULTADOS PRELIMINARES

As Tabelas 20, 21 e 22 apresentam os resultados para acurácia, *Exact Match* e *Hamming Loss*, sendo que os melhores resultados estão destacados em negrito.

Tabela 20 – Resultados do MultImp para a acurácia.

Base	MC	KNNI	MultiImp
birds	0,62667	0,60914	**0,62790**
CAL500	**0,43710**	0,42373	0,42655
emotions	**0,56799**	0,56790	**0,56799**
flags	0,71582	0,70614	**0,72093**

Até o momento da escrita desta tese, os resultados preliminares mostram que o método proposto é capaz de combinar de forma efetiva as soluções provenientes de métodos de imputação simples, otimizando sensivelmente as medidas de desempenho adotadas. Por meio da

Tabela 21 – Resultados do MultImp para o *exact match*.

Base	MC	KNNI	MultiImp
birds	0,56405	0,55706	**0,55714**
CAL500	0,21165	0,22371	**0,23368**
emotions	**0,45465**	0,45027	**0,45465**
flags	0,58456	0,58371	**0,59450**

Tabela 22 – Resultados do MultImp para o *Hamming Loss*.

Base	MC	KNNI	MultiImp
birds	**0,05075**	0,05352	0,052635
CAL500	**0,16323**	0,16753	0,16805
emotions	**0,25213**	0,25494	**0,25213**
flags	0,26537	0,27304	**0,26206**

análise dos resultados dispostos nas Tabelas 20, 21 e 22 é possível notar a eficiência na otimização dos parâmetros conforme a ordem lexicográfica imposta, pois o MultImp alcançou os melhores resultados para o EM e seu desempenho foi decaindo de acordo com os critérios de avaliação das funções objetivo.

7.3 IMPUTAÇÃO MÚLTIPLA PARA SÉRIES TEMPORAIS UTILIZANDO PROGRAMAÇÃO GENÉTICA

As principais motivações para tratar conjuntos de dados relacionados a séries temporais são a forte incidência de dados ausentes neste nicho de aplicação; e a capilaridade deste tipo de análise (*e.g.* matemática financeira, previsão meteorológica e geológica, telecomunicações). Como os conjuntos de dados destinados à analise de séries temporais são compostos quase que exclusivamente por atributos numéricos, a utilização de regressão para predizer os valores ausentes faz-se interessante (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

Neste contexto, a programação genética tem ganhado destaque, por suas habilidades de descobrir funções matemáticas que descrevam relações entre uma variável dependente e uma ou mais variáveis (independentes) de um conjunto de dados.

Conforme discutido no Capítulo 4, estatísticas baseadas na função de autocorrelação, média e variância são úteis para estimar a maior parte dos modelos lineares para séries temporais, como ARIMA e ARCH. Como apenas é possível calcular estas estatísticas a partir de exemplos sem valores ausentes, o GPImp utiliza-se de uma pré-imputação similar à abordagem utilizada por Tran, Zhang e Andreae (2015) - no método proposto, os valores ausentes são substituídos por valores aleatórios, para então iniciar o processo evolucionário de busca pela função de regressão. A subseção a seguir apresenta a estrutura genética e fluxo de trabalho do GPImp.

7.3.1 FUNCIONAMENTO DO GPIMP

No GPImp os indivíduos são árvores semânticas onde cada nó é uma função e as folhas são atributos independentes e constantes a fim de produzir a função de regressão de um atributo com valor ausente (f_i), o Algoritmo 3 apresenta o pseudocódigo do GPImp.

Algoritmo 3: Pseudocódigo do GPImp.

Entrada: Base de dados com valores ausentes (X_{mv}), parâmetros do GPImp
Saída : Base de dados imputada (X_{aug}), funções de regressão (Fc)

```
1  X_im ← base de dados com valores ausentes ignorados;
                                        /* F é o conjunto de atributos de X_mv */
2  para para cada atributo f_i de F faça
3      se o atributo f_i possui valores ausentes então
4          Inicializar os indivíduos;
5          Avaliar a população;
6          enquanto o número de gerações não for atingido ou fitness > erro faça
7              Aplicar operadores genéticos;
8              Avaliar os indivíduos;
9          fim
10         fc_i ← função de regressão obtida a partir do melhor indivíduo;
11     fim
12     para cada instância x ∈ X_mv faça
13         se f_i de x possui VA então
14             Usar a função de regressão fc_i para o atributo f_i na instância x;  /* O valor obtido pela
                função de regressão será o valor imputado */
15         fim
16         Imputar o valor obtido pelo passo anterior em X_aug e em X_mv ;
17     fim
18 fim
19 retorna Base Imputada (X_aug), funções de regressão (Fc);
```

O Algoritmo 3 recebe como entrada um conjunto de dados com valores ausentes juntamente com a parametrização inerente a um algoritmo de programação genética (i.e. tamanho da população, conjunto de funções e especificação dos operares). Como saída, a base imputada e o conjunto de funções de regressão para cada atributo com valor ausente (Fc) são retornados.

Definidas as entradas e saídas, parte-se para o pré-processamento do método, na Linha 1, define-se a base X_{im} como uma base onde as instâncias que apresentam valores ausentes são ignoradas. Então, para cada atributo do conjunto de dados de entrada, verifica-se se possuem valores ausentes (Linhas 2 e 3). Caso a condição da Linha 3 seja verdadeira, inicia-se a busca pela função de regressão para o atributo com valor ausente fc_i utilizando os demais atributos como argumentos da função (terminais). A heurística de busca adotada é a programação genética, onde primeiro gera-se a população inicial para posterior avaliação dos indivíduos (Linhas 4 e 5). Com a população inicial instanciada, parte-se para o processo evolucionário, aplicando sucessivamente os operadores genéticos de seleção, cruzamento e mutação, até atingir a condição de parada (Linhas 6-9).

A função de aptidão dos indivíduos é semelhante à utilizada por Figueroa García, Kalenatic e López Bello (2010) e visa minimizar a distância entre a função de autocorrelação, média

e variância do atributo f_i pertencente à X_{im} e o RMSE calculado a partir dos valores observados de f_i e os preditos pela fc_i, a partir da sub base sem valores ausentes (X_{im}). A seção a seguir descreve o cálculo da função objetivo.

A Linha 10 do Algoritmo 3 é responsável por retornar a função de regressão fc_i obtida para o atributo f_i. De posse dessa função de regressão, inicia-se o processo de imputação (Linhas 12-17). Primeiramente, identificam-se quais instâncias da base X_{mv} o atributo em questão possui valores ausentes. Para estas instâncias, utiliza-se fc_i para predizer o valor a ser imputado nas respectivas instâncias com base nos outros atributos - caso haja outros atributos com valores ausentes, considera-se o valor da média para predizer f_i - então imputa-se os valores preditos nas respectivas instâncias nas bases X_{mv} e X_{aug} (Linha 16). O fim do algoritmo ocorre na Linha 19, quando a base imputada (X_{aug}) e o conjunto contendo as funções de regressão (Fc) é retornado.

7.3.2 FUNÇÃO OBJETIVO ADOTADA NO GPIMP

Segundo Figueroa García, Kalenatic e López Bello (2010), séries temporais possuem propriedades diferenciadas que inviabilizam o uso de processos convencionais de análise de dados, conforme discutido no Capítulo 3. Por exemplo, uma série temporal possui estruturas autocorrelacionadas, podendo conter componentes ou tendências sazonais. Neste horizonte, três estatísticas são consideradas importantes ao se tratar valores ausentes de séries temporais: a média, a variância e as estruturas de autocorrelação da série temporal; considerando que estes são descritores completos do processo estocástico.

A autocovariância e a autocorrelação de uma série temporal são estatísticas utilizadas para estimar vários modelos como ARIMA, ARCH e GARCH, representando a distância do valor medido entre um tempo específico $\{f_{it}\}$ e sua defasagem h, $\{f_{i(t+h)}\}$, definido como a relação linear entre as duas medidas. A função de autocovariância amostral ($\hat{\gamma}(h)$) e a função de autocorrelação amostral ($\hat{\rho}(h)$) são calculadas de acordo com as Eq. 7.4 e 7.5, respectivamente.

$$\hat{\gamma}(h) = \sum_{t=1}^{n-|h|} \frac{(f_{i(t+|h|)} - \bar{f}_i)(f_{it} - \bar{f}_i)}{n}, \quad -n < h < n. \tag{7.4}$$

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n. \tag{7.5}$$

Onde i e n denotam o número de atributos e de amostras, respectivamente; e $\bar{f}_i$ é a média do atributo f_i. O objetivo da imputação é predizer valores que não alterem estas características, para isso, a função objetivo ($\mathcal{F}$) é calculada a partir da Eq. 7.6, onde H é um parâmetro de atraso h-*lag*.

$$\min \mathcal{F} = \sum_{h=1}^{H} |\hat{\rho}(h)^l - \hat{\rho}(h)| + |\bar{f}_i^a - \bar{f}_i| + |Var(X)^a - Var(X)| + RMSE \tag{7.6}$$

7.3.3 EXPERIMENTOS COMPUTACIONAIS

Para avaliar o desempenho do GPImp, seis conjuntos de dados de séries temporais foram utilizados, sendo dois obtidos a partir do repositório de aprendizado de máquina UCI (emg_lower_limb_apie_1, istanbul_stock) e quatro provenientes do repositório do KEEL (Acont 1_2000, Edat_1_1661, NN5_Complete_110 e NNGC1_D1_V1_002). Originalmente estes conjuntos de dados não possuem valores ausentes, por isso removeram-se valores de forma randômica, de forma obter 5, 10 e 30% de dados em falta. A Tabela 23 apresenta um resumo dos conjuntos de dados utilizados nos experimentos.

Tabela 23 – Bases de dados utilizadas

Nome	(%) VA	Nº Inst.	Nº Atrib.	(%) VA/Inst.	Total de VA
Acount_1_2000	5	1995	6	26,47	598
	10	1995	6	46,67	1197
	30	1995	6	89,97	3591
emg_lower_limb_apie_1	5	11403	5	22,62	2850
	10	11403	5	40,66	5701
	30	11403	5	83,66	17104
Edat_1_1661	5	1655	6	26,59	496
	10	1655	6	47,67	993
	30	1655	6	89,55	2979
istanbul_stock	5	536	9	36,19	241
	10	536	9	61,94	482
	30	536	9	95,15	1447
NN5_Complete_110	5	787	6	26,94	236
	10	787	6	47,65	472
	30	787	6	89,83	1416
NNGC1_D1_V1_002	5	1175	6	27,23	352
	10	1175	6	47,49	705
	30	1175	6	88,77	2115

O desempenho do GPImp foi comparado em relação a três métodos de imputação comumente utilizados neste nicho de aplicação, a saber: *Event Covering*; imputação K-*Means*; e imputação por média. Duas medidas de avaliação foram utilizadas, o NRMSE é usado para avaliar a distância entre os valores imputados e os valores reais; e o coeficiente de correlação obtido pelo algoritmo SMOReg (máquina de vetor de suporte para regressão) (SHEVADE et al., 2000), este algoritmo foi escolhido pois alcança bons resultados e vem se destacando na área. Utilizaram-se os parâmetros padrões para o SMOReg, e para particionar o conjunto de dados em treino e teste utilizou-se a validação cruzada com 10-*folds*.

A parametrização do GPImp seguiu os indicativos propostos por Tran, Zhang e Andreae (2015) e Figueroa García, Kalenatic e López Bello (2010). Do primeiro autor, importaram-se os parâmetros característicos de um método baseado em programação genética, como conjunto de primitivos e atributos qualitativos, como os operadores de inicialização, cruzamento e mutação. Do Segundo, adaptaram-se os parâmetros quantitativos comuns aos métodos evolucionários,

como taxa de mutação, taxa de cruzamento; e também a taxa de defasagem (*h-lag*).

Quadro 7.1 – Parâmetros utilizados no GPImp.

Parâmetro	Valor
Conjunto de Funções	$+$, $-$, $\times$, $\backslash$, *exp*, *sin*, *cos*, *abs*
Terminais Variáveis	todos os atributos exceto o atributo de interesse
Terminais Constantes	valores aleatórios
Inicialização	*Hamped half-and-half*
Cruzamento	*Subtree crossover*
Mutação	*Subtree mutation*
Seleção	Torneio
Tamanho da População	1024
Número de Gerações	50
Número de indivíduos por torneio	7
Taxa de Cruzamento	60%
Taxa de Mutação	30%
Taxa de Reprodução	10%
h-lag	7

7.3.3.1 RESULTADOS PARA O RMSE

A Tabela 24 apresenta os valores do NRMSE obtidos por cada método de imputação para os conjuntos de dados analisados; onde quanto maior o valor do NRMSE, maior a proximidade dos valores imputados para com os valores reais. Ao final da tabela é apresentado o ranqueamento dos métodos obtidos por meio da aplicação do teste de Friedman.

Conforme visto na Tabela 24, o GPImp foi suplantado apenas pelo KMI; no entanto, não encontrou-se evidência estatística da superioridade, conforme atestado pelos testes *post-hoc* Holm e Shaffer, apresentados na Tabela 25 para um intervalo de confiança de 95%.

Por meio da análise dos resultados dispostos na Tabela 25 é possível concluir que o KMI e o GPImp são estatisticamente superiores ao MC e ao EC, mas equivalentes entre si.

7.3.3.2 RESULTADOS PARA AS ESTATÍSTICAS

A Tabela 26 apresenta os resultados para as diferenças estatísticas, considerando a média, variância e função de autocorrelação entre o conjunto de dados pré-imputado e a base imputada. Para a média da função de aptidão para todos os atributos retirando-se o argumento do RMSE - quanto menor a diferença, melhor o resultado.

Os melhores resultados da Tabela 26 estão marcados em negrito, como se pode observar, o GPImp mostrou-se competitivo frente aos outros métodos. Este fato é evidenciado pelos resultados do coeficiente de correlação obtido por meio da aplicação do SMOReg - por definição, o último atributo foi considerado o atributo a ser predito. O coeficiente de correlação mede a força e a direção da relação linear entre variáveis, os valores são sempre no intervalo de $+1$ a

Tabela 24 – Resultados do GPImp para o NRMSE.

Base	EC	GPImp	KMI	MC
Acount_1_2000	0,586	1	0,907	0
Acount_1_2000	0,447	0	1	0,168
Acount_1_2000	0,545	0,451	1	0
Edat_1_1661	0,162	1	0,354	0
Edat_1_1661	0,433	1	0,545	0
Edat_1_1661	0,447	1	0,712	0
emg_lower_limb_apie_1	0,79	1	0,984	0
emg_lower_limb_apie_1	0,881	0,919	1	0
emg_lower_limb_apie_1	0,88	0,148	1	0
istanbul_stock	0	0,58	1	0,468
istanbul_stock	0	1 0,967	0,915	
istanbul_stock	0	0,125	1	0,265
NN5_Complete_110	0,443	1	0,838	0
NN5_Complete_110	0,356	1	0,837	0
NN5_Complete_110	0,269	0,966	1	0
NNGC1_D1_V1_002	0,194	1	0,805	0
NNGC1_D1_V1_002	0	0,991	1	0,173
NNGC1_D1_V1_002	0	0,824	1	0,243
Rank	**3.112**	**1.778**	**1.5**	**3.612**

Tabela 25 – *p*-valores ajustados pelos métodos de Holm e Shaffer para intervalo de confiança de 95%.

Hipóteses	*p*-valor	Holm	Shaffer
KMI vs, MC	0,00009E-9	0,00833	0,00833
GPImp vs, MC	0,00002	0,01	0,01667
EC vs, KMI	0,00018	0,0125	0,01667
EC vs GPImp	0,00195	0,01667	0,01667
EC vs MC	0,24528	0,025	0,025
GPImp vs KMI	0,51861	0,05	0,05

-1, quando próximos de $|1|$, denotam uma forte relação linear, enquanto mais próximos de 0, menor é a relação.

A Tabela 27 mostra os valores do coeficiente de correlação, com os valores mais distantes de zero grafados em negrito. Por meio da análise dos resultados é possível perceber que o conjunto de dados imputado pelo GPImp apresenta bons resultados para o coeficiente de correlação. Além do desempenho, chama-se atenção para a interpretabilidade do método proposto, uma vez que o GPImp fornece uma função de regressão, como mostra a Figura 27.

Os operadores presentes na função de regressão apresentada na Figura 27, *abs*, *sin* e *cos* denotam valor absoluto, seno e cosseno, respectivamente. Os nós-folha são as variáveis terminais e constantes, no caso, apenas os atributos F[1], F[3] e F[4]; F[0] - *timestamp* e F[2] não foram considerados neste indivíduo. Uma hipótese para este fato é que estes atributos não

Tabela 26 – Resultados para as diferenças estatísticas.

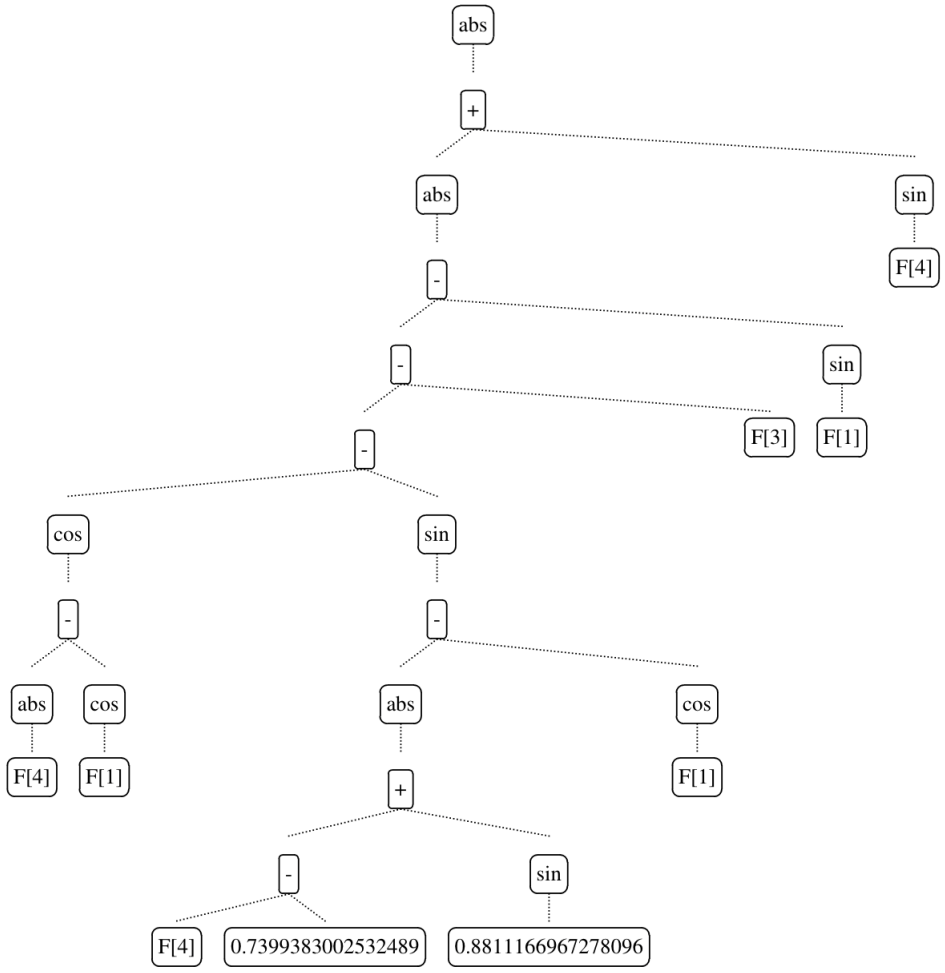
Nome	(%) VA	EC	GPImp	KMI	MC
Acount_1_2000	5	**7,121**	15,052	79,668	39,229
	10	37,816	**8,253**	117,884	72,852
	30	**36,927**	118,635	473,525	240,101
Edat_1_1661	5	0,03	**0,006**	0,026	0,035
	10	0,044	**0,006**	0,056	0,073
	30	0,148	**0,013**	0,112	0,176
emg_lower_limb_apie_1	5	**11,282**	57,948	13,831	30,258
	10	**20,328**	116,898	30,425	50,366
	30	41,914	**27,891**	47,47	44,624
istanbul_stock	5	0,007	**0,003**	**0,003**	**0,003**
	10	0,014	**0,006**	**0,006**	0,009
	30	0,032	0,016	**0,009**	0,012
NN5_Complete_110	5	1,334	**1,059**	2,464	1,897
	10	5,101	**1,188**	6,657	3,229
	30	21,548	**6,137**	16,14	6,303
NNGC1_D1_V1_002	5	**63,979**	74,051	733,322	586,313
	10	**427,296**	442,403	2332,528	632,986
	30	**1970,039**	3343,989	3704,605	8086,619

Tabela 27 – Coeficiente de correlação obtido pelo SMOreg.

Base	(%) VA	EC	GPImp	KMI	MC
Acount_1_2000	5	0,78627	**0,808613**	0,78967	0,76448
	10	**0,71532**	0,687853	0,73860	0,66928
	30	0,63399	**0,804569**	0,57280	0,32548
Edat_1_1661	5	0,79773	**0,842489**	0,79716	0,78896
	10	0,76518	**0,842186**	0,74726	0,72270
	30	0,61222	**0,869179**	0,65300	0,55888
emg_lower_limb_apie_1	5	5,24E-4	0,020208	**0,02584**	0,01214
	10	**-0,02270**	-0,004358	-0,01343	0,00444
	30	0,00866	7,312E-4	**0,12369**	-0,03740
istanbul_stock	5	0,82803	0,855859	**0,86648**	0,85469
	10	0,64463	**0,853378**	0,84049	0,81881
	30	0,41615	0,377949	**0,77879**	0,6859
NN5_Complete_110	5	0,660815	**0,70341**	0,68934	0,64973
	10	0,627165	**0,718599**	0,66059	0,56535
	30	0,511349	**0,866686**	0,66038	0,38466
NNGC1_D1_V1_002	5	0,781842	**0,813468**	0,81003	0,76676
	10	0,710552	**0,811869**	0,80925	0,70483
	30	0,507051	**0,874724**	0,85449	0,53254

trazem ganho para a predição do rótulo.

Figura 27 – Função de regressão obtida para o atributo “V1/-3” do conjunto de dados “NN5”.



Fonte: Elaborada pelo autor.

7.4 CONSIDERAÇÕES FINAIS

Neste capítulo, dois métodos de imputação de dados foram apresentados, o MultImp e o GPImp. O primeiro método apresentado, o MultImp, teve dois objetivos principais que guiaram sua concepção: i) preencher uma lacuna encontrada na literatura no que tange à não existência de uma análise do impacto dos dados ausentes no contexto da classificação multirrótulo; ii) testar o conceito de evolução de métodos de imputação, onde utiliza-se métodos de imputação simples como indivíduos da população inicial para então aplicar o processo evolucionário para combiná-los. Este processo iterativo confere ao método as vantagens da imputação múltipla discutidas anteriormente. Com o intuito de compensar um possível tamanho de população diminuto, o MultImp é dotado de uma estratégia para controle deste parâmetro em tempo de execução que, embora simples, apresentou bons resultados, superando os métodos de imputação comparados e reduzindo o custo computacional de métodos como o GAImp e MOGAImp.

O segundo, baseado em programação genética, utiliza estatísticas como média, variância

e função de autocorrelação para compor a função de aptidão que guia o processo evolucionário. Como resultado, o GPImp provê um os dados imputados e um conjunto de funções de regressão para cada um dos atributos da série temporal. Este diferencial confere ao método uma capacidade de interpretabilidade, possibilitando um melhor entendimento do padrão de ausência de dados e facilitando a incorporação de conhecimento de fundo por meio da modificação das funções de regressão. Ademais, os resultados obtidos pelo GPImp mostraram-se promissores, elegendo-o como uma alternativa viável para utilização no tratamento de valores ausentes em séries temporais.

8 CONCLUSÕES

Neste capítulo são apresentadas as conclusões deste trabalho por meio da retomada das perguntas de pesquisa definidas na introdução, correlacionando-as com as contribuições científicas alcançadas por meio de publicações em conferências e periódicos. Também são discutidas as potencialidades e limitações dos métodos propostos neste trabalho, sugerindo direcionamentos para trabalhos futuros e destacando as dificuldades encontradas.

8.1 AVALIAÇÃO DAS PERGUNTAS DE PESQUISA

No capítulo 1 foram definidos três objetivos principais que tratam dos pontos de pesquisa em aberto e que pertencem ao escopo desta tese. A seguir, as perguntas de pesquisa que nortearam o desenvolvimento do trabalho em cada um dos objetivos são discutidas, apresentando as soluções obtidas.

Objetivo 1: Qual a importância de um modelo formal para imputação múltipla de dados como um problema de otimização? Como definir e representar o espaço de busca e restrições de forma a respeitar as características intrínsecas da base? Quais as estratégias de busca que melhor se aplicam ao modelo formal proposto?

A utilização de uma descrição da imputação múltipla de dados como um problema de otimização propicia um melhor entendimento do problema sob seus diferentes aspectos, possibilitando a proposição de métodos de tratamento de valores ausentes aos mais variados cenários até então não explorados; bem como a adaptação de métodos existentes, a fim de se moldarem a problemas específicos como classificação desbalanceada, classificação multirrótulo e *big data*, por exemplo.

Com uma descrição também é possível: i) utilizar diferentes representações para o espaço de busca a fim de reduzir o custo computacional do processo de busca de valores a serem imputados - tal como as diferenças nas representações utilizadas no AGImp e no MOGAImp (Capítulos 5 e 6) que, embora sejam sutis, impactam diretamente no custo computacional; ii) adaptar estratégias de busca, como no caso do MultImp (Capítulo 7), onde utilizam-se soluções provenientes de métodos de imputação simples, a fim de combiná-las e refiná-las por meio de uma estratégia evolucionária, tornando o método iterativo e de imputação múltipla; iii) identificar automaticamente estratégias de busca de acordo com as características intrínsecas da base e dos algoritmos de análise a serem utilizados (*e.g.* correlacionar o algoritmo de classificação a um método de imputação baseado em otimização combinatorial).

Portanto, é possível afirmar que objetivo desta tese, a saber: "*Propor e testar uma definição formal para a imputação múltipla de dados como um problema de otimização, permitindo*

suplantar as falhas presentes nos métodos de imputação de dados baseados em CE recentemente propostos.", foi parcialmente atingido, uma vez que mesmo com a sua potencialidade, foi apenas proposta uma descrição do problema em detrimento ao modelo formal. Descrição esta que necessita de aprimoramentos, sobretudo no estudo das múltiplas representações do espaço de busca.

Objetivo 2: Quais as vantagens e desvantagens da utilização de computação evolucionária para realizar a imputação de dados? Como tratar atributos numéricos e categóricos igualitariamente e ainda levar em consideração exemplos com valores ausentes? Como avaliar soluções candidatas e combinar as soluções levando-se em consideração medidas de desempenho conflitantes?

A principal vantagem da utilização da computação evolucionária para realização da imputação de dados dá-se pela sua proximidade com o paradigma da imputação múltipla. Dentre as vantagens desta categoria, destacam-se: a produção de estimativas imparciais, proporcionando maior robustez do que abordagens *ad hoc*; utilização de todos os dados disponíveis, preservando o tamanho da amostra; e ainda ser possível a utilização do software estatístico (ou de aprendizado de máquina) já utilizado pelos analistas. Tais vantagens são decorrentes da sua modularidade e estratégia de estimação dos valores ausentes e combinação das diversas soluções produzidas.

A principal desvantagem é o custo computacional associado à iteratividade do método e o impacto da inicialização no processo de busca, por conseguinte, estratégias que diminuam o espaço de busca ou que guiem o processo de inicialização se fazem interessantes, tal como abordado nos métodos GAImp e MultImp propostos nos Capítulos 5 e 7, respectivamente.

Em relação ao tratamento de atributos numéricos e categóricos igualitariamente, sua discretização e ordenamento são estratégias interessantes, há também a possibilidade de se utilizar aproximadores de funções de densidade de probabilidade e faixas de valores. Tais estratégias possibilitam a utilização de exemplos que possuam valores ausentes na imputação de dados, não excluindo informações potencialmente úteis do processo de análise.

No tocante à avaliação de soluções candidatas, considerando medidas de desempenho conflitantes, é possível empreender diversas abordagens de otimização multiobjetivo. A abordagem baseada em fronteira de Pareto é útil para o estudo da relação de dominância entre as medidas de desempenho consideradas, apesar de seu custo computacional mais alto; já a baseada em lexicografia é uma alternativa para evitar o custo computacional da exploração do conjunto de Pareto. Tais abordagens foram utilizadas no MOGAImp e no MultImp, respectivamente apresentadas nos Capítulos 6 e 7.

Frente ao exposto, é possível asseverar que o objetivo "*Desenvolver e aprimorar algoritmos de imputação múltipla de dados baseado em CE eficientes, que considerem conjuntos de dados com atributos de tipos mistos, evitem a análise de caso completo e que lidem de forma*

satisfatória com medidas de desempenho conflituosas." foi plenamente alcançado.

Objetivo 3: Qual o impacto da variação dos parâmetros no desempenho do método? As informações acerca da convergência e parametrização são importantes para o domínio de aplicação ou podem ajudar o especialista do domínio a melhor entender a ausência de dados na base em análise?

Conforme evidenciado na análise dos experimentos computacionais desenvolvidos (Seção 5.4; Seção 6.3), a variação dos parâmetros influenciam diretamente no desempenho e no custo computacional do método, por conseguinte, fez-se necessário um estudo acerca da sintonização dos parâmetros. Como conclusões, tem-se: i) para o GAImp, o tamanho da população e o número de gerações são os parâmetros mais sensíveis - observou-se que, ao utilizar um número de indivíduos superior ao número de gerações, obtém-se um melhor custo-benefício entre desempenho e custo computacional; ii) para o MOGAImp, é necessário um maior número de indivíduos e de gerações que o GAImp, tanto pelo seu esquema de codificação, quanto pela utilização de múltiplas funções objetivo - percebeu-se também que quanto mais valores ausentes um conjunto de dados apresenta, mais indivíduos e gerações são necessárias para garantir um bom desempenho. Estratégias de inicialização e de redução do espaço de busca fazem-se interessantes, uma vez que a iteratividade do método aumenta o custo computacional associado, este decorrente da construção de modelos de classificação.

Outro item pertinente é que as informações acerca da convergência e da parametrização podem auxiliar ao especialista no estudo do conjunto de dados, principalmente no tocante ao padrão de valores ausentes. Por exemplo, a adoção de múltiplas funções objetivo permitem ao especialista melhor entender as relações entre propriedades estatísticas da base, os padrões de ausência e o impacto na construção do modelo de análise (*e.g* classificação, regressão, agrupamento) - até mesmo, provendo informações no projeto de novos experimentos, como adição ou remoção de variáveis.

Nesse horizonte, é possível considerar que objetivo "*Analisar o comportamento de métodos de imputação múltipla baseados em computação evolucionária em relação à convergência e parametrização; de forma a estudar estratégias de sintonização e controle de parâmetros adequadas.*" foi alcançado.

8.2 RESUMO DAS PRODUÇÕES

As contribuições deste trabalho podem ser consideradas em dois aspectos: i) apuração da literatura e aspectos teóricos dos experimentos envolvendo imputação de dados; ii) proposição de métodos de imputação baseados em computação evolucionária e de um modelo formal para a imputação de dados como um problema de otimização combinatorial. A seguir, as produções são listadas e classificadas de acordo sua divulgação: publicadas, em prelo e em preparação para submissão.

No que concerne ao primeiro aspecto, são consideradas as seguintes contribuições:

- Revisão sistemática sobre métodos de tratamento de valores ausentes: uma revisão sistemática no tema foi planejada, conduzida e inicialmente reportada neste projeto de tese. Em sua etapa de condução, 9.000 publicações foram identificadas, com 132 artigos passando pelos critérios de seleção e 40 trabalhos devidamente analisados. Como resultado, percebeu-se uma tendência clara no uso de imputação de dados como o principal método para lidar com VA, adicionalmente, percebeu-se uma falta de padronização nos experimentos, o que dificulta a replicação, avaliação e comparação fidedigna entre os métodos recentemente propostos, seja pela academia ou indústria. Tal revisão foi submetida a um periódico onde os revisores sugeriram a realização de melhorias no estilo de escrita e uma análise mais aprofundada dos métodos selecionados para análise. **Divulgação:** em preparação para ressubmissão.

- *Framework* experimental para testes envolvendo imputação de dados: foi proposto um *framework* para a implementação de testes envolvendo métodos de imputação de dados no contexto do aprendizado supervisionado. A motivação para o desenvolvimento deste *framework* foi identificada na revisão sistemática conduzida: a falta de padronização nos testes envolvendo métodos para tratamento de valores ausentes, o que dificulta a replicação e consequentemente a comparação fidedigna entre eles. Portanto, o objetivo deste trabalho é fornecer aos pesquisadores uma sequência de etapas que permitam a replicação dos experimentos no contexto de classificação de padrões. **Divulgação:** publicado em conferência internacional.

- Revisão acerca dos métodos de imputação baseados em computação evolucionária: foi realizada uma análise dos trabalhos que utilizam computação evolucionária no processo de imputação de dados. O objetivo deste trabalho é levantar oportunidades e desafios que permeiam a área a fim de guiar estudos futuros. **Divulgação:** em preparação para submissão.

No que concerne ao segundo aspecto, é possível destacar:

- Algoritmo genético para a imputação múltipla de dados (GAImp): foi proposto e desenvolvido um algoritmo genético mono-objetivo para a imputação de dados para otimizar classificadores baseados em aprendizado de máquina. A maior parte dos métodos de imputação são restritos a um tipo de variável apenas (categóricas ou numéricas) e recaem em análise de caso completo. Portanto, o método proposto visa preencher tais lacunas na literatura, lidando de forma satisfatória com os tipos de dados supracitados, além de levar em consideração instâncias com valores ausentes. Como função de avaliação, adotou-se a acurácia do classificador, de forma a incorporar a informação da construção do modelo

na escolha dos valores a serem imputados. Os resultados mostraram-se promissores e corroboram os métodos de imputação baseados em algoritmos genéticos como heurística para implementar a imputação múltipla de dados. **Divulgação:** publicado em conferência internacional.

- Extensão dos resultados do algoritmo genético multiobjetivo: em um trabalho posterior, o método foi adaptado e os testes foram estendidos a conjuntos de dados adicionais, tanto com VA existentes, quanto com induzidos. Também avaliou-se a convergência das soluções e a sensibilidade da parametrização do algoritmo genético para a imputação de dados. Os resultados mostram que o método proposto obtém performance superior aos métodos de imputação comparados; e o comportamento do algoritmo genético desenvolvido é estudado em relação à adoção de diferentes valores para os parâmetros quantitativos. Neste trabalho, uma versão preliminar do modelo formal da imputação de dados como um problema de otimização combinatorial foi introduzido. **Divulgação:** em prelo (periódico) - trabalho submetido a uma edição especial de Otimização combinatória aplicada - em um primeiro momento solicitou-se uma revisão a qual foi submetida e aguarda-se o posicionamento dos revisores.

- Algoritmo genético multiobjetivo para imputação múltipla de dados (MOGAImp): foi proposto e desenvolvido um algoritmo genético multiobjetivo para a imputação de dados. O algoritmo é baseado no algoritmo NSGA-II e incorpora as características do algoritmo mono-objetivo proposto, levando em consideração instâncias com VA e informação da construção do modelo de classificação, e ainda, lida com atributos mistos da mesma forma. Os resultados obtidos mostram que o método multiobjetivo proposto apresenta um bom custo-benefício para medidas de avaliação conflitantes, o método mostra-se flexível quanto ao domínio de aplicação, uma vez que a função de avaliação pode ser facilmente modificada. **Divulgação:** publicado em periódico.

- Método de imputação de dados para séries temporais utilizando programação genética (GPImp): foi proposto e desenvolvido um método baseado em programação genética para identificar funções de regressão, as quais são utilizadas para predizer os valores a serem imputados, visando manter as características estatísticas do conjuntos de dados. Um dos diferenciais deste método concerne à sua interpretabilidade, uma vez que as funções de regressão podem ser facilmente entendidas pelos analistas. Os resultados mostraram-se satisfatórios mas são necessárias modificações para tornar o método competitivo, dentre elas, vislumbrou-se a adoção de uma estratégia para diminuição do espaço de busca por agrupamento. **Divulgação:** publicado em conferências nacionais.

- Método de imputação de dados para classificação multirótulo (MultImp): foi proposto e desenvolvido um método de imputação múltipla para classificação multirótulo que baseia-se na ideia de combinação de soluções obtidas por métodos de imputação simples. Isto

feito por meio de um processo evolucionário a fim de obter uma solução mais robusta ao final do processo. Para compor a função objetivo, três medidas de desempenho de classificação multirrótulo foram utilizadas em uma abordagem multiobjetivo lexicográfica, a acurácia do classificador, o casamento exato e o *Hamming Loss*. Os resultados mostraram-se promissores, e a estratégia de inicialização e controle de população em tempo de execução diminuem o seu custo computacional, tornando o MultImp um método de imputação atrativo a este nicho de aplicação.

Destaca-se também a elaboração deste documento de tese, que contém a descrição dos estudos conduzidos, métodos propostos e análises realizadas. Por fim, conclui-se que o trabalho desenvolvido traz benefícios para os diversos domínios onde a análise de dados com dados ausentes se faz notória, sejam acadêmicos ou industriais, uma vez que os métodos propostos são escaláveis e adaptáveis, sobretudo na classificação de padrões e na análise de séries temporais.

8.3 OBSTÁCULOS DE PESQUISA ENCONTRADOS E TRABALHOS FUTUROS

Ao longo do desenvolvimento da tese, alguns obstáculos referentes à complexidade da pesquisa foram encontrados. Os mais relevantes, são:

- A ubiquidade e capilaridade da ausência de dados faz com que estudos para mitigar seus efeitos danosos sejam conduzidos nos mais variados campos. Sendo assim, observa-se na literatura uma grande quantidade de trabalhos desenvolvidos para este fim, distribuídos nas mais diversas áreas. É possível encontrar a proposição e análise de métodos de imputação em periódicos/conferências em análise de dados, inteligência computacional, redes de computadores, sensoriamento remoto, epidemiologia, bioinformática e estatística, por exemplo. À vista disso, encontrou-se uma dificuldade no tocante à busca, seleção, filtragem, comparação e análise dos referidos estudos;
- Também pela ubiquidade dos valores ausentes na análise de dados, há uma série de formalismos, sejam eles matemáticos ou estatísticos, que fundamentam o processo. No entanto, observa-se na literatura o uso de diferentes notações para a descrição de um mesmo fenômeno. Outro ponto pertinente é que nem sempre a descrição de determinados métodos é clara ou possui implementação publicamente disponível. Tais fatos, dificultaram a compreensão dos formalismos matemáticos/estatísticos que fundamentam a análise de dados com valores ausentes, bem como abordagens de tratamento recentemente propostas;
- Apesar da ausência de dados ser um problema frequente, encontrar estudos de casos reais, onde a quantidade de valores ausentes não inviabiliza a análise de novos algoritmos, e que seja possível incorporar conhecimento de fundo dos analistas, é uma dificuldade a ser pontuada.

A partir das contribuições e dificuldades elencadas, é possível identificar algumas limitações nos métodos propostos e nas análises realizadas - tais limitações dão margem para desdobramento dos estudos devido à própria natureza do processo de desenvolvimento de um trabalho científico. Dessa forma, alguns direcionamentos podem ser apontados como sugestões de trabalhos futuros, tais como:

- Foram propostas e analisadas diversas estratégias para codificação e inicialização das soluções candidatas; além de diversas medidas de desempenho a serem utilizadas como funções objetivo. Um possível desdobramento desta tese diz respeito à análise de diferentes combinações entre as estratégias propostas, avaliando o impacto no desempenho do método e no custo computacional envolvido;
- Tais estratégias também são passíveis de modificações e de adaptações, incluindo a utilização de outras heurísticas de busca como otimização por colônia de formiga ou outros algoritmos multiobjetivo como o SPEA; a análise de diferentes estratégias de inicialização, incluindo abordagens que utilizem-se da função de densidade aproximada de cada atributo ou que implementem restrições do tipo *Must-Link* e *Cannot-Link*, e extrair regras de imputação a partir das soluções fornecidas pelos métodos desenvolvidos, por exemplo;
- Um dos fatos da não adoção de métodos de imputação em estudos de casos reais mais recorrentemente apontados na literatura é a dificuldade em parametrização do método. Portanto, estudar alternativas para redução do número de parâmetros, ou ainda, desenvolver uma abordagem adaptativa que leve em consideração as características da ausência de dados e dos algoritmos de análise é um ponto passível de ser utilizado em trabalhos futuros.

Evidentemente, as sugestões de trabalhos futuros acima mencionadas podem ser combinadas de diferentes maneiras, e ainda, avaliadas em diferentes cenários.

REFERÊNCIAS

ABDELLA, M.; MARWALA, T. The use of genetic algorithms and Neural Networks to approximate missing data in database. *Computing and Informatics*, v. 24, p. 577–589, 2005. Citado na página 53.

ABDELLA, M.; MARWALA, T. Treatment of missing data using neural networks and genetic algorithms. In: *IEEE International Joint Conference on Neural Networks*. [S.l.]: Ieee, 2005. v. 2, p. 598–603. Citado na página 53.

ACUNA, E.; RODRIGUEZ, C. The Treatment of Missing Values and its Effect on Classifier Accuracy. In: BANKS, D. et al. (Ed.). *Classification, Clustering, and Data Mining Applications SE - 60*. [S.l.]: Springer Berlin Heidelberg, 2004. p. 639–647. Citado 2 vezes nas páginas 48 e 53.

AITTOKALLIO, T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, v. 11, n. 2, p. 253–264, mar. 2010. Citado na página 52.

ALBA, E.; LUQUE, G.; NESMACHNOW, S. Parallel metaheuristics: Recent advances and new trends. *International Transactions in Operational Research*, v. 20, n. 1, p. 1–48, 2013. Citado na página 1.

ALCALÁ, J. et al. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, v. 17, n. 2-3, p. 255–287, 2010. Citado 2 vezes nas páginas XIII e 88.

ALLISON, P. D. *Missing Data (Quantitative Applications in the Social Sciences)*. [S.l.]: Sage publications, 2001. Citado 2 vezes nas páginas 4 e 14.

ALVARES-CHERMAN, E.; METZ, J.; MONARD, M. C. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, v. 39, n. 2, p. 1647–1655, fev. 2012. Citado 2 vezes nas páginas 103 e 104.

ANDRIDGE, R. R.; LITTLE, R. J. A. A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, v. 78, n. 1, p. 40–64, abr. 2010. Citado 2 vezes nas páginas 41 e 42.

ARMELLINI, F.; KAMINSKI, P. C.; BEAUDRY, C. Integrating open innovation to new product development - the case of the Brazilian aerospace industry. *Int J of Technological Learning Innovation and Development*, v. 5, n. 4, p. 367–384, 2012. Citado na página 10.

AYDILEK, I. B.; ARSLAN, A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, v. 233, p. 25–35, jun. 2013. Citado na página 3.

BÄCK, T. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford, UK: Oxford University Press, 1996. Citado 2 vezes nas páginas 1 e 22.

BANZHAF, W. et al. *Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and Its Applications*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. Citado 2 vezes nas páginas 22 e 33.

BARALDI, A. N.; ENDERS, C. K. An introduction to modern missing data analyses. *Journal of school psychology*, v. 48, n. 1, p. 5–37, fev. 2010. Citado 2 vezes nas páginas 2 e 41.

BARROS, R.; BASGALUPP, M.; CARVALHO, A. de. Investigating fitness functions for a hyper-heuristic evolutionary algorithm in the context of balanced and imbalanced data classification. *Genetic Programming and Evolvable Machines*, v. 16, n. 3, p. 241–281, 2015. Citado 2 vezes nas páginas 42 e 70.

BATISTA, G. E. A. P. A.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 2003. Citado 3 vezes nas páginas 40, 44 e 48.

BRABHAM, D. C. Crowdsourcing as a Model for Problem Solving. *Convergence: The International Journal of Research into New Media Technologies*, v. 14, n. 1, p. 75, 2008. Citado na página 10.

BRÁS, L. P.; MENEZES, J. C. Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular Engineering*, v. 24, n. 2, p. 273–282, 2007. Citado 2 vezes nas páginas 51 e 54.

BROWN, M. L.; KROS, J. F. Data mining: Opportunities and Challenges. In: *Data Mining: Opportunities and Challenges*. Hershey, PA, USA: IGI Publishing, 2003. Citado na página 14.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, v. 45, n. 1, p. 1–67, 2011. Citado 3 vezes nas páginas 52, 62 e 108.

CHANG, G.; GE, T. Comparison of missing data imputation methods for traffic flow. In: *International Conference on Transportation, Mechanical, and Electrical Engineering*. [S.l.]: Ieee, 2011. p. 639–642. Citado na página 52.

CHEEMA, J. R. A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, v. 84, n. 4, p. 487–508, 2014. Citado na página 46.

CHEN, H. Y.; LITTLE, R. A test of missing completely at random for generalized. *Biometrika*, v. 86, n. 1, p. 1–13, 1999. Citado na página 17.

CISMONDI, F. et al. Computational intelligence methods for processing misaligned, unevenly sampled time series containing missing data. *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, p. 224–231, abr. 2011. Citado na página 44.

Coello Coello, C. Multi-objective Evolutionary Algorithms in Real-World Applications: Some Recent Results and Current Challenges. In: GREINER, D. et al. (Ed.). *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences SE - 1*. [S.l.]: Springer International Publishing, 2015, (Computational Methods in Applied Sciences, v. 36). p. 3–18. Citado na página 31.

COVER, T. M.; THOMAS, J. A. *Elements of Information Theory*. [S.l.]: Wiley, 1991. (Wiley Series in Telecommunications, Wiley Series in Telecommunications). Citado na página 49.

de Andrade Silva, J.; HRUSCHKA, E. R. EACImpute: An Evolutionary Algorithm for Clustering-Based Imputation. *International Conference on Intelligent Systems Design and Applications*, p. 1400–1406, 2009. Citado 5 vezes nas páginas 3, 41, 53, 58 e 67.

De Castro, L. N. L. N. Fundamentals of natural computing: an overview. *Physics of Life Reviews*, v. 4, n. 1, p. 1–36, mar. 2007. Citado 5 vezes nas páginas 1, 21, 22, 24 e 37.

DEB, K.; KALYANMOY, D. *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2001. Citado na página 30.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Trans. Evol. Comp*, v. 6, n. 2, p. 182–197, abr. 2002. Citado 2 vezes nas páginas 31 e 81.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, v. 39, n. 1, p. 1–38, 1977. Citado na página 13.

DERRAC, J. et al. Analyzing convergence performance of evolutionary algorithms: A statistical approach. *Information Sciences*, v. 289, p. 41–58, dez. 2014. Citado na página 3.

DERRAC, J. et al. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, v. 1, n. 1, p. 3–18, mar. 2011. Citado 2 vezes nas páginas 72 e 89.

DING, Y.; ROSS, A. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, v. 45, n. 3, p. 919–933, mar. 2012. Citado na página 52.

DOQUIRE, G.; VERLEYSEN, M. Feature selection with missing data using mutual information estimators. *Neurocomputing*, v. 90, p. 3–11, ago. 2012. Citado na página 51.

DORRI, F.; AZMI, P.; DORRI, F. Missing value imputation in DNA microarrays based on conjugate gradient method. *Computers in biology and medicine*, v. 42, n. 2, p. 222–7, fev. 2012. Citado na página 42.

EEKHOUT, I. et al. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*, v. 23, n. 5, 2012. Citado 3 vezes nas páginas 2, 3 e 46.

EIBEN, A.; SMIT, S. Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm and Evolutionary Computation*, v. 1, n. 1, p. 19–31, mar. 2011. Citado 4 vezes nas páginas 4, 26, 27 e 29.

EIBEN, A. E.; HINTERDING, R.; MICHALEWICZ, Z. Parameter control in evolutionary algorithms. *Evolutionary Computation, IEEE Transactions on*, v. 3, n. 2, p. 124–141, jul. 1999. Citado na página 28.

EIBEN, A. E.; JELASITY, M. A critical note on experimental research methodology in EC. In: *Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on*. [S.l.: s.n.], 2002. v. 1, p. 582–587. Citado na página 29.

EIBEN, A. E.; SCHIPPERS, C. A. On evolutionary exploration and exploitation. *Fundamenta Informaticae*, v. 35, n. 1-4, p. 35–50, 1998. Citado na página 23.

EIBEN, A. E.; SMITH, J. From evolutionary computation to the evolution of things. *Nature*, v. 521, n. 7553, p. 476–482, maio 2015. Citado 4 vezes nas páginas 1, 2, 22 e 24.

EIBEN, A. E.; SMITH, J. E. *Introduction to Evolutionary Computing*. [S.l.]: SpringerVerlag, 2003. Citado 3 vezes nas páginas 1, 23 e 29.

ENDERS, C. K. *Applied Missing Data Analysis*. [S.l.]: Guilford Press, 2010. (Methodology in the social sciences). Citado na página 47.

ENDERS, C. K. Analyzing longitudinal data with missing values. *Rehabilitation psychology*, v. 56, n. 4, p. 267–88, nov. 2011. Citado na página 44.

ESPEJO, P. G.; VENTURA, S.; HERRERA, F. A Survey on the Application of Genetic Programming to Classification. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, v. 40, n. 2, p. 121–144, 2010. Citado na página 33.

FACELI, K. et al. *Inteligência Artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LCT, 2011. Citado 3 vezes nas páginas 13, 37 e 49.

FARHANGFAR, A.; KURGAN, L.; DY, J. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, v. 41, n. 12, p. 3692–3705, dez. 2008. Citado 3 vezes nas páginas 42, 44 e 48.

FÁVERO, L. P. et al. *Análise Multivariada de Dados: Modelagem multivariada para tomada de decisões*. [S.l.]: CAMPUS, 2009. Citado 2 vezes nas páginas 1 e 11.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, p. 37–54, 1996. Citado 2 vezes nas páginas 11 e 37.

FERRO, M. a. Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of epidemiology*, v. 24, n. 1, p. 75–7, jan. 2014. Citado na página 44.

Figueroa García, J.; KALENATIC, D.; López Bello, C. Missing Data Imputation in Time Series by Evolutionary Algorithms. In: HUANG, D.-S. et al. (Ed.). *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence SE - 34*. [S.l.]: Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 5227). p. 275–283. Citado 4 vezes nas páginas 3, 55, 57 e 58.

Figueroa García, J. C.; KALENATIC, D.; López Bello, C. A. An Evolutionary Approach for Imputing Missing Data in Time Series. *Journal of Circuits, Systems and Computers*, v. 19, n. 01, p. 107–121, feb. 2010. Citado 9 vezes nas páginas 3, 4, 42, 56, 58, 70, 109, 110 e 111.

Figueroa García, J. C.; KALENATIC, D.; López Bello, C. A. Missing data imputation in multivariate data by evolutionary algorithms. *Computers in Human Behavior*, v. 27, n. 5, p. 1468–1474, sep. 2011. Citado 4 vezes nas páginas 3, 40, 56 e 58.

FLORES, P.; COTA, M. G.; MORALES, L. B. Modeling Time series with missing and incorrect values using Self Adaptive Genetic Algorithms. In: *Int'l Conf. Genetic and Evolutionary Methods*. [S.l.: s.n.], 2011. p. 175–180. Citado na página 42.

FOGEL, L. J. *Intelligence Through Simulated Evolution: Forty Years of Evolutionary Programming*. New York, NY, USA: John Wiley & Sons, Inc., 1999. Citado na página 21.

FRANÇA, F. de; COELHO, G.; Von Zuben, F. Predicting missing values with biclustering: A coherence-based approach. *Pattern Recognition*, v. 46, n. 5, p. 1255–1266, maio 2013. Citado na página 54.

FRANÇA, F. O. D. *Biclusterização na Análise de Dados Incertos*. 198 p. Tese (Doutorado) — Universidade Estadual de Campinas, 2010. Citado 3 vezes nas páginas 54, 55 e 58.

FRANKLIN, J. The formal sciences discover the philosophers stone. *Studies in History and Philosophy of Science Part A*, v. 25, n. 4, p. 513–533, ago. 1994. Citado na página 38.

FREITAS, A. A. A Critical Review of Multi-objective Optimization in Data Mining: A Position Paper. *SIGKDD Explor. Newsl.*, New York, NY, USA, v. 6, n. 2, p. 77–86, 2004. Citado na página 31.

GARCÍA-LAENCINA, P. J.; SANCHO-GóMEZ, J.-L.; FIGUEIRAS-VIDAL, A. R. Pattern classification with missing data: a review. *Neural Computing and Applications*, v. 19, n. 2, p. 263–282, set. 2009. Citado 11 vezes nas páginas 1, 2, 3, 4, 13, 14, 20, 37, 44, 45 e 46.

GARCÍA-LAENCINA, P. J. et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, v. 72, n. 7-9, p. 1483–1493, mar. 2009. Citado na página 51.

GARCÍA, S.; LUENGO, J.; HERRERA, F. Dealing with Missing Values. In: *Data Preprocessing in Data Mining SE - 4*. [S.l.]: Springer International Publishing, 2015, (Intelligent Systems Reference Library, v. 72). p. 59–105. Citado na página 1.

GAUTAM, C.; RAVI, V. Evolving clustering based data imputation. In: *Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on*. [S.l.: s.n.], 2014. p. 1763–1769. Citado 6 vezes nas páginas 42, 55, 56, 57, 58 e 67.

GAUTAM, C.; RAVI, V. Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, 2015. Citado 2 vezes nas páginas 42 e 57.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. [S.l.]: Addison-Wesley, 1989. (Artificial Intelligence, Addison-We). Citado na página 22.

GÓMEZ-CARRACEDO, M. et al. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, v. 134, p. 23–33, maio 2014. Citado na página 53.

GONCALVES, E. C.; PLASTINO, A.; FREITAS, A. A. A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In: *Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence*. Washington, DC, USA: IEEE Computer Society, 2013. (ICTAI '13), p. 469–476. Citado 3 vezes nas páginas 70, 104 e 107.

GONcALVES, E. C.; PLASTINO, A.; FREITAS, A. A. Simpler is better: A novel genetic algorithm to induce compact multi-label chain classifiers. In: *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*. [S.l.: s.n.], 2015. (GECCO '15), p. 559–566. ISBN 978-1-4503-3472-3. Citado na página 107.

GRAHAM, J. W. Missing data analysis: making it work in the real world. *Annual review of psychology*, v. 60, p. 549–76, jan. 2009. Citado 8 vezes nas páginas 1, 3, 13, 14, 16, 37, 44 e 46.

GROSSMAN, G. M.; HELPMAN, E. Outsourcing in a Global Economy. *The Review of Economic Studies*, v. 72, n. 1, p. 135–159, 2005. Citado na página 10.

GRZYMALA-BUSSE, J. W.; HU, M. A comparison of several approaches to missing attribute values in data mining. In: *Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*. London, UK, UK: Springer-Verlag, 2001. (RSCTC '00), p. 378–385. Citado 2 vezes nas páginas 41 e 48.

GUO, C.-x. et al. Swarm intelligence for mixed-variable design optimization. *Journal of Zhejiang University Science*, v. 5, n. 7, p. 851–860, 2004. Citado na página 38.

HAIR, J. F. *Multivariate data analysis*. Seventh edition. [S.l.]: Harlow, Essex : Prentice Hall, 2014. Citado 2 vezes nas páginas 1 e 10.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2006. (The Morgan Kaufmann series in data management systems, Second Edition). Citado 5 vezes nas páginas 1, 10, 11, 12 e 37.

HECKMAN, J. J. Sample Selection Bias as a Specification Error. *Econometrica*, v. 47, n. 1, p. 153–161, 1979. Citado na página 13.

HEERINGA, S.; WEST, B.; BERGLUND, P. *Applied survey data analysis*. [S.l.]: Chapman & Hall/CRC, 2010. Citado na página 13.

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems*. Cambridge, MA, USA: MIT Press, 1992. Citado 2 vezes nas páginas 21 e 22.

HONAKER, J.; KING, G. G.; KING, G. G. What about Missing Data to Do Values in Time-Series. *American Journal of Political Science*, v. 54, n. 2, p. 561–581, 2013. Citado 3 vezes nas páginas 4, 44 e 103.

HOWE, J. The Rise of Crowdsourcing. *North*, v. 14, n. 14, p. 1–5, 2006. Citado na página 10.

HRUSCHKA, E.; Hruschka Jr., E.; EBECKEN, N. Towards Efficient Imputation by Nearest-Neighbors: A Clustering-Based Approach. In: WEBB, G.; YU, X. (Ed.). *AI 2004: Advances in Artificial Intelligence SE - 45*. [S.l.]: Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, v. 3339). p. 513–525. Citado na página 54.

HRUSCHKA, E. R. et al. On the influence of imputation in classification: practical issues. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 21, n. 1, p. 43–58, mar. 2009. Citado 4 vezes nas páginas 3, 42, 44 e 57.

HRYDZIUSZKO, O.; VIANT, M. R. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, v. 8, n. S1, p. 161–174, out. 2011. Citado na página 44.

HUNG, J.-C. A genetic algorithm approach to the spectral estimation of time series with noise and missed observations. *Information Sciences*, v. 178, n. 24, p. 4632–4643, dez. 2008. Citado 2 vezes nas páginas 42 e 44.

INMAN, D.; ELMORE, R.; BUSH, B. A case study to examine the imputation of missing data to improve clustering analysis of building electrical demand. *Building Services Engineering Research and Technology*, v. 36, n. 5, p. 628–637, set. 2015. Citado na página 53.

JESUS, L. D. et al. A Testbed for the Experiments Performed in Missing Value Treatments. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, v. 55, n. 91, 2013. Citado na página 7.

JUNGER, W.; Ponce de Leon, A. Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, v. 102, p. 96–104, fev. 2015. Citado 2 vezes nas páginas 4 e 44.

KANG, S.-S.; KOEHLER, K.; LARSEN, M. D. *Partial FEFI for incomplete tables with covariates*. Valley View, Ames, 2007. Citado na página 21.

KARAFOTIAS, G.; HOOGENDOORN, M.; EIBEN, A. E. Parameter Control in Evolutionary Algorithms: Trends and Challenges. *Evolutionary Computation, IEEE Transactions on*, v. 19, n. 2, p. 167–187, abr. 2015. Citado 3 vezes nas páginas 4, 28 e 29.

KENNEDY, J.; EBERHART, R.; SHI, Y. *Swarm intelligence*. [S.l.]: The Journal of the American College of Dentists, 2001. Citado na página 22.

KIM, J. K. Parametric fractional imputation for missing data analysis. *Biometrika*, v. 98, n. 1, p. 119–132, mar. 2011. Citado na página 21.

KIM, K.-Y.; KIM, B.-J.; YI, G.-S. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, v. 5, n. 1, p. 160, 2004. Citado na página 54.

KONAK, A.; COIT, D. W.; SMITH, A. E. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, v. 91, n. 9, p. 992–1007, set. 2006. Citado 2 vezes nas páginas 30 e 31.

KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992. Citado na página 34.

KRISHNA, M.; RAVI, V. Particle swarm optimization and covariance matrix based data imputation. In: *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*. [S.l.: s.n.], 2013. p. 1–6. Citado 4 vezes nas páginas 42, 56, 57 e 58.

KWAK, N.; CHOI, C.-H. C. C.-H. Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 12, p. 1667–1671, 2002. Citado 2 vezes nas páginas 49 e 51.

LIAO, T. et al. Ant Colony Optimization for Mixed-Variable Optimization Problems. *Evolutionary Computation, IEEE Transactions on*, v. 18, n. 4, p. 503–518, 2014. Citado na página 38.

LICHMAN, M. *UCI Machine Learning Repository*. 2013. Disponível em: <http://archive.ics-.uci.edu/ml>. Citado 5 vezes nas páginas XIII, 61, 71, 88 e 89.

LIEW, A. W.-C.; LAW, N.-F.; YAN, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, v. 12, n. 5, p. 498–513, set. 2011. Citado na página 52.

LITTLE, R. J. A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, v. 83, n. 404, p. 1198–1202, 1988. Citado na página 17.

LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: Wiley, 1987. 278 p. Citado 5 vezes nas páginas 4, 13, 15, 37 e 78.

LITTLE, R. J. A.; RUBIN, D. B. *Statistical Analysis with missing data*. 2. ed. New York: Wiley, 2002. Citado 7 vezes nas páginas 1, 2, 3, 4, 13, 15 e 62.

LIU, Y.; BROWN, S. D. Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Intelligent Laboratory Systems*, v. 120, p. 106–115, jan. 2013. Citado 4 vezes nas páginas 51, 53, 57 e 58.

LOBATO, F. et al. Multi-Objective Genetic Algorithm For Missing Data Imputation. *Pattern Recognition Letters*, set. 2015. Citado 2 vezes nas páginas 7 e 102.

LOBATO, F. M. F. et al. An Evolutionary Missing Data Imputation Method for Pattern Classification. In: *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*. New York, NY, USA: ACM, 2015. (GECCO Companion '15), p. 1013–1019. Citado 3 vezes nas páginas 7, 41 e 80.

LUENGO, J.; GARCÍA, S.; HERRERA, F. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, v. 32, n. 1, p. 77–108, jun. 2012. Citado 10 vezes nas páginas 1, 49, 51, 53, 54, 60, 70, 72, 76 e 85.

LUENGO, J.; SÁEZ, J. a.; HERRERA, F. Missing data imputation for fuzzy rule-based classification systems. *Soft Computing*, v. 16, n. 5, p. 863–881, out. 2011. Citado na página 51.

LUKE, S. *Essentials of Metaheuristics*. second. [S.l.]: Lulu, 2013. Citado na página 1.

MARGARIA, T. Service Is in the Eyes of the Beholder. *Computer*, v. 40, n. 11, p. 33–37, 2007. Citado na página 10.

MCCLEARY, L. Using Multiple Imputation for Analysis of Incomplete Data in Clinical Research. *Nursing Research*, v. 51, n. 5, 2002. Citado na página 61.

MCKNIGHT, P. et al. *Missing Data: A Gentle Introduction (Methodology In The Social Sciences)*. [S.l.]: The Guilford Press, 2007. Citado 5 vezes nas páginas 16, 17, 18, 44 e 47.

MIETTINEN, K. *Nonlinear Multiobjective Optimization*. [S.l.]: Springer US, 1999. (International Series in Operations Research & Management Science). Citado na página 30.

MUKHOPADHYAY, A. et al. A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I. *Evolutionary Computation, IEEE Transactions on*, v. 18, n. 1, p. 4–19, 2014. Citado na página 32.

MUKHOPADHYAY, A. et al. Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II. *Evolutionary Computation, IEEE Transactions on*, v. 18, n. 1, p. 20–35, 2014. Citado na página 32.

MUNDFROM, D. J. et al. *Imputing Missing Values: The Effect on the Accuracy of Classification*. 1998. Citado na página 48.

NANNI, L.; LUMINI, A.; BRAHNAM, S. A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, v. 55, n. 1, p. 37–50, maio 2012. Citado na página 44.

NEWMAN, D. a. Missing Data: Five Practical Guidelines. *Organizational Research Methods*, v. 17, n. 4, p. 372–411, set. 2014. Citado 2 vezes nas páginas 10 e 44.

NIELSEN, M. A.; CHUANG, I. L. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. 10th. ed. New York, NY, USA: Cambridge University Press, 2011. Citado na página 21.

OLIVEIRA, P. de; COELHO, A. Genetic Versus Nearest-Neighbor Imputation of Missing Attribute Values for RBF Networks. In: KÖPPEN, M.; KASABOV, N.; COGHILL, G. (Ed.). *Advances in Neuro-Information Processing SE - 34*. [S.l.]: Springer Berlin Heidelberg, 2009, (Lecture Notes in Computer Science, v. 5507). p. 276–283. Citado na página 1.

OLIVEIRA, P. G. de. *Imputação automática de atributos faltantes em problemas de classificação: um estudo comparativo envolvendo algoritmos bio-inspirados*. Tese (Doutorado) — Universidade de Fortaleza, 2009. Citado 2 vezes nas páginas 13 e 37.

PAN, R. et al. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. *Applied Intelligence*, v. 43, n. 3, p. 614–632, 2015. Citado na página 51.

PATIL, D. V.; BICHKAR, R. S. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques. In: . [S.l.: s.n.], 2010. p. 74–78. Citado 3 vezes nas páginas 40, 57 e 58.

PEREIRA, J. C. R. *Análise de datos qualitativos: estrategias metodológicas para as ciencias da saúde, humanas o sociais*. [S.l.]: EDUSP, 2004. Citado na página 11.

PEYRE, H.; LEPLÈGE, A.; COSTE, J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French. *Quality of life research*, v. 20, n. 2, p. 287–300, mar. 2011. Citado na página 52.

POLI, R.; LANGDON, W. B.; MCPHEE, N. F. *A Field Guide to Genetic Programming*. [S.l.]: Lulu Enterprises, UK Ltd, 2008. Citado 3 vezes nas páginas 33, 35 e 36.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Citado na página 43.

PRICE, K.; STORN, R. M.; LAMPINEN, J. A. *Differential Evolution: A Practical Approach to Global Optimization (Natural Computing Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005. Citado na página 22.

QUINLAN, J. R. Unknown attribute values in induction. In: *Proceedings of the Sixth International Workshop on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989. p. 164–168. Citado na página 46.

RAO, S. S.; XIONG, Y. A Hybrid Genetic Algorithm for Mixed-Discrete Design Optimization. *Journal of Mechanical Design*, v. 127, n. 6, p. 1100–1112, out. 2004. Citado na página 38.

READ, J. et al. Classifier chains for multi-label classification. *Machine Learning*, v. 85, n. 3, p. 333–359, jun. 2011. Citado na página 104.

REZENDE, S. O. et al. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Manole Ltda, 2003. Citado 2 vezes nas páginas 11 e 12.

ROTHLAUF, F. *Design of Modern Heuristics: principles and Application*. [S.l.]: Springer Berlin Heidelberg, 2011. (Natural Computing Series). Citado 3 vezes nas páginas 1, 3 e 22.

RUBIN, D. B. Inference and missing data. *Biometrika*, v. 63, n. 3, p. 581–592, 1976. Citado na página 13.

RUBIN, D. B. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons., 1987. Citado 4 vezes nas páginas 21, 61, 62 e 63.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. Citado na página 21.

SAINANI, K. L. Dealing With Missing Data. *PM&R*, v. 7, n. 9, p. 990–994, set. 2015. Citado na página 1.

SCHAFER, J. L. *Analysis of Incomplete Multivariate Data*. [S.l.]: Chapman & Hall, 1997. 164–165 p. (C&H/CRC Monographs on Statistics & Applied Probability, 3). Citado 2 vezes nas páginas 2 e 4.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. *Psychological Methods*, v. 7, n. 2, p. 147–177, 2002. Citado 7 vezes nas páginas 16, 18, 37, 44, 46, 57 e 61.

SCHAFFER, J. D. Multiple objective optimization with vector evaluated genetic algorithms. In: *Proceedings of the 1st International Conference on Genetic Algorithms*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1985. p. 93–100. Citado na página 31.

SCHUTT, R.; O'NEIL, C. *Doing Data Science: Straight Talk from the Frontline*. [S.l.]: O'Reilly Media, Inc., 2013. Citado na página 1.

SCHWEFEL, H.-P. *Numerical Optimization of Computer Models*. New York, NY, USA: John Wiley & Sons, Inc., 1981. Citado na página 21.

SHEVADE, S. K. et al. Improvements to the smo algorithm for svm regression. *Trans. Neur. Netw.*, v. 11, n. 5, p. 1188–1193, set. 2000. Citado na página 111.

SILVA, J. D. A. *Substituição de Valores ausentes: uma abordagem baseado em um algoritmo evolutivo para agrupamento de dados*. Tese (Doutorado) — Universidade de São Paulo, 2010. Citado 2 vezes nas páginas 13 e 37.

SILVA, J. D. A.; HRUSCHKA, E. R. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data & Knowledge Engineering*, v. 84, p. 47–58, jan. 2013. Citado 3 vezes nas páginas 3, 51 e 53.

SIM, J.; KWON, O.; LEE, K. C. Adaptive Pairing of Classifier and Imputation Methods Based on the Characteristics of Missing Values in Data Sets. *Expert Systems with Applications*, nov. 2015. Citado 2 vezes nas páginas 52 e 85.

SONG, Q.; KASABOV, N. Ecm - a novel on-line, evolving clustering method and its applications. In: *In M. I. Posner (Ed.), Foundations of cognitive science*. [S.l.]: The MIT Press, 2001. p. 631–682. Citado na página 55.

SOUTO, M. C. P. de; JASKOWIAK, P. A.; COSTA, I. G. Impact of missing data imputation methods on gene expression clustering and classification. *BMC bioinformatics*, v. 16, n. 1, p. 64, jan. 2015. Citado 2 vezes nas páginas 52 e 53.

SRINIVAS, N.; DEB, K. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evol. Comput.*, v. 2, n. 3, p. 221–248, set. 1994. Citado na página 31.

TANG, J. et al. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transportation Research Part C: Emerging Technologies*, v. 51, p. 29–40, fev. 2015. Citado 2 vezes nas páginas 55 e 58.

TRAN, C. T.; ANDREAE, P.; ZHANG, M. Impact of imputation of missing values on genetic programming based multiple feature construction for classification. In: *Evolutionary Computation (CEC), 2015 IEEE Congress on*. [S.l.: s.n.], 2015. p. 2398–2405. Citado 2 vezes nas páginas 44 e 52.

TRAN, C. T.; ZHANG, M.; ANDREAE, P. Multiple Imputation for Missing Data Using Genetic Programming. In: *Proceedings of the 2015 on Genetic and Evolutionary Computation Conference*. New York, NY, USA: ACM, 2015. (GECCO '15), p. 583–590. Citado 5 vezes nas páginas 41, 42, 58, 108 e 111.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, v. 17, n. 6, p. 520–525, jun. 2001. Citado na página 54.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Mining multi-label data. In: *In Data Mining and Knowledge Discovery Handbook*. [S.l.: s.n.], 2010. p. 667–685. Citado na página 104.

TSOUMAKAS, G. et al. Mulan: A java library for multi-label learning. *J. Mach. Learn. Res.*, v. 12, p. 2411–2414, jul. 2011. Citado na página 107.

Van Hulse, J.; KHOSHGOFTAAR, T. M. Incomplete-case nearest neighbor imputation in software measurement data. *Information Sciences*, jan. 2011. Citado na página 2.

VERONEZE, R. Assessing the Performance of a Swarm-based Biclustering Technique for Data Imputation. In: *IEEE Congress on Evolutionary Computation*. [S.l.: s.n.], 2011. p. 386–393. Citado na página 54.

VERONEZE, R. *Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla*. Tese (Doutorado) — Universidade Estadual de Campinas, 2011. Citado 4 vezes nas páginas 13, 37, 54 e 55.

WAGSTAFF, K. et al. Constrained k-means clustering with background knowledge. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 577–584. Citado 2 vezes nas páginas 2 e 41.

WILCOXON, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80–83, 1945. Citado na página 49.

WILSON, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *Ieee Transactions On Systems Man And Cybernetics*, v. 2, n. 3, p. 408–421, 1972. Citado na página 49.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. [S.l.]: Morgan Kaufmann, 2011. (Morgan Kaufmann series in data management systems, 2). Citado 2 vezes nas páginas 1 e 12.

WOHLRAB, L.; FÜRNKRANZ, J. A review and comparison of strategies for handling missing values in separate-and-conquer rule learning. *Journal of Intelligent Information Systems*, v. 36, n. 1, p. 73–98, abr. 2010. Citado na página 46.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1–37, 2008. Citado na página 12.

XIE, H. Analyzing longitudinal clinical trial data with nonignorable missingness and unknown missingness reasons. *Computational Statistics & Data Analysis*, v. 56, n. 5, p. 1287–1300, maio 2012. Citado na página 44.

YOUNG, R.; JOHNSON, D. R. Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. *Journal of Marriage and Family*, v. 77, n. 1, p. 277–294, fev. 2015. Citado na página 44.

YOZGATLIGIL, C. et al. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology*, v. 112, n. 1-2, p. 143–167, 2013. Citado na página 52.

ZHANG, S. Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, v. 35, n. 1, p. 123–133, fev. 2010. Citado 2 vezes nas páginas 20 e 63.

ZHANG, S. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, v. 85, n. 11, p. 2541–2552, nov. 2012. Citado na página 2.

ZHANG, S.; JIN, Z.; ZHU, X. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, v. 23, n. 3, p. 110–121, mar. 2011. Citado na página 2.

ZHI-XIN, W.; JU, G. A parallel genetic algorithm in multi-objective optimization. In: *Control and Decision Conference, 2009. CCDC '09. Chinese*. [S.l.: s.n.], 2009. p. 3497–3501. Citado na página 31.

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, n. 1, mar. 2011. Citado 2 vezes nas páginas 30 e 31.

ZITZLER, E.; LAUMANNS, M.; THIELE, L. *SPEA2: Improving the Strength Pareto Evolutionary Algorithm*. [S.l.], 2001. Citado na página 31.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *Evolutionary Computation, IEEE Transactions on*, v. 3, n. 4, p. 257–271, nov. 1999. Citado na página 31.

...

ANEXOS

ANEXO A – TRABALHOS PUBLICADOS E PROJETO DE PESQUISA.

Periódicos:

Lobato, F. M. F. ; Sales Junior, C. S. ; Araujo, I. M. ; Tadaiesky, V. W. A. ; Dias, L. J. C. ; Ramos, L. ; Santana, Á. L. Multi-objective genetic algorithm for missing data imputation. Pattern Recognition Letters, v. 68, p. 126-131, 2015.

Conferências Internacionais:

Lobato, F. M. F. ; Tadaiesky, V. W. A. ; Araujo, I. M. ; Santana, Á. L. . An Evolutionary Missing Data Imputation Method for Pattern Classification. In: Genetic and Evolutionary Computation Conference, 2015, Madrid. Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference. New York, NY, USA: ACM, 2015. p. 1013-1019.

Dias, L. J. C. ; **Lobato, F. M. F.** ; Santana, Á. L. . A Testbed for the Experiments Performed in the Missing Value Treatments. In: International Conference on Information Society, 2013, Veneza. Proceedings of the International Conference on Information Society, 2013.

Conferências Nacionais:

Lobato, F. M. F.; Resende, D. C. O. ; Santana, Á. L. . Time Series Imputation Using Genetic Programming And Lagrange Interpolation. In: Brazilian Conference On Intelligent System, 2016, Recife, Pe. 5th Brazilian Conference On Intelligent System, 2016.

Resende, D. C. O. ; Santana, Á. L. ; Jacob Junior, A. F. L. **; Lobato, F. M. F.** . Multivariate Time Series Imputation Using Genetic Programming. In: Simpósio Brasileiro De Pesquisa Operacional, 2016, Vitória, Es. Anais Do XlVIII Sbpo, 2016.

Projeto de Pesquisa:

Título: Desenvolvimento de modelos bioinspirados aplicados ao tratamento de valores ausentes;

Descrição: Este projeto visa investigar a utilização de modelos bioinspirados ao tratamento de valores ausentes com o intuito de se mitigar os efeitos danosos desta problemática ubíqua nos mais diversos âmbitos, tais como: classificação de padrões, regressão e análise de séries temporais.

Local: Universidade Federal do Oeste do Pará

Período: 2014 - 2016

ANEXO B – REVISÃO SISTEMÁTICA

Missing Values Treatment: A Systematic Review

Fábio Manoel França Lobato [§12], Vincent Willian Araújo Tadaiesky[1], Lilian de Jesus Chaves Dias[1], Antônio Fernando Lavareda Jacob Junior[12], Ádamo Lima de Santana[1]

[1]Technological Institute, Federal University of Pará (UFPa), Belém, PA, Brazil.
[2] Science and Technology Centre, University of Amazon (UNAMA), Belém, PA, Brazil.
[§]Corresponding author

FMFL: lobato.fabio@ufpa.br
VWAT: vincent@ufpa.br
LJCD: lilianchavesdias@gmail.com
AFLJJ: jacobjr@ufpa.br
ALS: adamo@ufpa.br

Abstract

Data analysis pass through many application domains, such as social costumer relationship management, document classification, medical diagnosis or routing tracking. Despite the data source, is probable to occur a ubiquitous problem in data analysis, the missing values. Many strategies to handle with this problem have been proposed, some approaches based on machine learning, other methods are imported from statistical learning theory. Even though the importance of missing data, there is a lack of a systematic literature review for it, thus, this paper aim to fill this gap and provides a systematic review about missing value treatment. Initially, more than 9.000 publications were analyzed and 40 journal papers were reviewed. As a result, it was perceived a clear trend in the use of data imputation as a preferred method to handle missing values. In addition, it was also found a lack of standardization in the experiments involving missing values it represents an obstacle on the evaluation, replication and comparison of recently proposed methods, which prevents the full usage of the methods by the academy or industry. Finally, it is expected that the prospects of this survey are to provide important insights for possible researches on the treatment of missing values.

Keywords: missing value treatment, missing data, data imputation, systematic review.

1. Introduction

It is clear the great increase in the amount of data produced in recent years. IBM® estimates that every day 2.5 quintillion bytes of data are created

(Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012), which led to the concept of Big Data. This massive amount of data brings along challenges regarding the storage, retrieval of information and data analysis process (Motavalli, 2012). Thus, in order to extract valuable knowledge from data, it is necessary to develop data analysis methods to provide competitive advantages to corporations.

In this context, the process of Knowledge Discovery in Database - KDD (Fayyad, Piatetsky-shapiro, & Smyth, 1996; Han & Kamber, 2006) can be applied for data analysis and knowledge acquisition. The objective of KDD is to identify behaviors and trends in different areas of knowledge. Fayyad et al. (1996) conceptualizes KDD as a multi-step, non-trivial, interactive and iterative process to identify understandable, valid, new and potentially useful patterns from databases.

However, a recurring problem in data analysis is data incompleteness – when there is missing information in certain instances (Little & Rubin, 2002). Consequently, it can seriously degrade the data analysis process, given that the techniques developed do not handle missing data. Therefore, it is essential to treat this problem in order to improve the quality of result. Missing data occur for many reasons, but flaws in the process of data acquisition and applications of surveys are the main causes.

Missing data is an old and high-capillarity problem, since it reaches every research niche requiring data collection and analysis. There is a large number of studies developed in the area, besides the diverse forms of data treatment aiming to reduce the bias imposed by the problem. The most investigated topics are treatment methods based on Machine Learning, along with the development and adaptation of robust inference methods to missing values (MV).

Given this scenario, this study provides a systematic review about MV in order to identify trends and opportunities for future researches. It was decided to use the systematic review, since it provides a rigorous methodological approach to grant reliable studies comparison and to facilitate the replication process. Consequently, there will be a replicable methodology for identifying, evaluating and interpreting all available and relevant researches on a

particular question, knowledge area or phenomenon (Kitchenham & Charters, 2007).

The main objective of this review is to perform a general analysis of current researches about missing data. From the studies analyzed, it was noticed a tendency to treat MV by means of data imputation. Furthermore, it was also identified a lack of standardization in the experimental methods, which brings a negative impact, once it prevents reliable comparison between the recently proposed methods.

This paper is organized as follows: Section 2 presents a brief theoretical background about MV; Section 3 presents the conduction of a systematic review of missing value treatment (MVT); Section 4 presents the evaluation and discussion of the selected papers; and Section 5 presents the conclusions.

2. Theoretical Background

In a dataset, the absence of information is called Missing Data. Other terms are also used such as missing values and incomplete data (Little & Rubin, 2002). This problem is considered ubiquitous in data analysis process (Heeringa, West & Berglund, 2010) and may have different causes. Brown and Kros (2003) illustrate a few categories related to the MV sources:

- Operational factors: data entry errors, incorrect estimation, and accidental removal of table fields;
- Refusal to answer research questions;
- Impossibility of applying a particular questionnaire.

The operational factors are more common in the KDD context. Representative examples are errors in data input on information systems; problems in data warehousing step; communication network failure; and data collection device malfunction. The causes mentioned have a direct influence in the problem treatment, Little and Rubin (2002) describe the mechanisms of missingness according to their randomness as follows:

- Missing completely at random (MCAR): a situation that occurs when the probability of missing data on the variable is independent of the variable itself or any other influence;
- Missing at random (MAR): the missingness of data is independent of the missing values, but the absence pattern can be predicted by other variables in the database;
- Not missing at random (NMAR): the pattern of missing data is not random and it depends on the missing value itself.

In the MCAR mechanism, the missingness of a variable does not depend on the input values. The available examples contain all the information to make inferences. Typical examples of the MCAR mechanism are tubes containing a blood sample that accidentally break, resulting in the inability to measure blood parameters. The cause of data loss is completely random and the probability of missing an observation is not related to any other characteristic of the individual.

On the other hand, in the MAR mechanism, the missingness of a variable depends only on the values observed in the input dataset. For example, the occasional failure of a sensor due to a power outage, which interrupts the acquisition process. In this example, the current variables, where data is missing, are not the cause of incompleteness because the absence was caused by an external influence.

In contrast to the MAR pattern, the missingness of a variable in the NMAR mechanism cannot be predicted considering only the dataset variables. For example, if a sensor is not able to get information out of a certain range, the data missing is because of the NMAR. Then, it is said that the data was censored and consequently important information is lost, thus there is not an appropriate method to handle this type of missingness mechanism.

As for the MCAR or MAR patterns, there is a consensus to call them as ignorable patterns, in other words, such mechanisms are easy to handle since their effects on statistical models and currently in machine learning methods are available to analysts (Schafer, 1997; Mcknight, Mcknight, Sidani, & Figueredo, 2007; Graham, 2009). Therefore, studies addressing the MVT consider the MAR as the major missingness mechanism.

Focusing on the studies involving the MVT, García-Laencina, Sancho-Gómez, and Figueiras-Vidal (2009) divide them into four classes:

- Traditional approaches: also called complete case analysis. The missing data is treated by simple omission of instances or attributes with missing values, examples are list wise deletion and pair wise deletion;
- Imputation: replaces the values associated to the missing data, usually null or "?", for a plausible value. They are subdivided into statistical and machine learning methods, the first applies statistical measures to estimate the values to be imputed, meanwhile the second method uses machine learning algorithms to predict MV. Examples are single imputation, multiple imputation, and k-nearest neighbor (KNN) imputation;
- Models: aims to apply methods of maximum-likelihood estimation, in order to determine the joint distribution of each attribute and find the value to be imputed. Although the methods based on models perform data imputation, a considerable number of authors do not include them in the imputation class. Examples of this category are Expectation-Maximization (EM) algorithm and Gaussian Mixture Models;
- Machine Learning Methods: avoid explicit imputation, targeting the development/adaptation of the machine learning algorithms in order to improve their robustness regarding the incidence of missing data. Methods based on ensemble classifiers and fuzzy methods illustrate some techniques.

Regardless of the chosen approach, the goal is to reduce the bias imposed by missing data that inherently affect the data analysis result, considering that the analysis techniques were not modeled to directly handle MV. The concepts presented about mechanisms and MVT classes allowed the design and conduction of this systematic review, which is described in the following section.

3. Conducting the Systematic Review

As previously mentioned, the decision to apply the Systematic Review in the missing values treatment was made due to the volume of recent publications on the subject. Therefore, starting from the need to apply this research

methodology, a pilot review to identify keywords on the topic was developed. The following research questions were prepared:

1. What are the papers involving treatment of missing values?
2. Which papers propose and compare imputation methods?
3. What studies embrace a standardized proposal for experiments to treat missing values?

In order to answer these questions, the following search terms were selected:

a) Missing Data
b) Missing Values
c) Incomplete Data
d) Data Imputation
e) Missing Data Imputation
f) Missing Data Treatment
g) Multiple Imputations
h) Multivariate Analysis
i) Machine Learning
j) Pattern Classification
k) Testbed
l) Protocol
m) Framework

The terms from a) to h) represent the main subject of this research, while i) and j) terms provide refinements to the search query. During the research, it was noticed a lack of standardization of MVT experiments. Therefore, it was included a third research question and another refinement using terms from k) to m).

It is possible to combine search terms in periodical databases using logical operators such as AND and OR, an example of combination used in this

systematic review is "Missing Data Treatment" AND "Data imputation" AND "Testbed". These search terms were applied to databases accessible from the CAPES Journal Portal[1], which collects multiple bases such as:

1. Elsevier Science Direct;
2. ACM Digital Library Biomedical Central;
3. Cambridge University Press;
4. Free E-Journals;
5. Institute of Physics;
6. Nature;
7. Oxford University Press;
8. Scielo;
9. Willey On-Line Library.

Other databases with no correlation to the subject, either directly or indirectly, were omitted from this list. In order to filter the large amount of results, a selection criteria similar to the Spolaôr, Cherman, Metz, and Monard (2013) was defined, as shown in Table 1.

Identifier	Selection Criteria
01	Publications not related to Missing Values Treatment (such as missing values prevention)
02	Duplicate papers with the same authors and abstract.
03	Studies performing pre-processing.
04	Papers addressing the robustness of machine learning algorithms according to missing values.
05	Publications lacking new methodologies (comparison or application of methods previously proposed)
06	Papers consisting of only one page such as abstracts; posters; presentations and seminars.

[1] www.periodicos.capes.gov.br/

07	Papers hosted on databases inaccessible from CAPES portal.
08	Papers lacking experimental results report or review of other publication from the literature
09	Papers written in languages other than English.
10	Books, chapters of books, completion of coursework papers, dissertations and theses.
11	Publications outside the last triennium.

Table 1: List of selection criteria used in this systematic review.

It is important to note that the selection criteria exclude studies of subsequent analysis while the quality criteria measure the methodological quality of a study. Spolaôr et al. (2013) present one contribution to these criteria, which is the possibility of correlating results of different publications according to the analysis quality. For instance, the use of statistical tests to evaluate the hypotheses. Table 02 lists the quality criteria used in this systematic review

Identifier	Quality Criteria
01	Does the study compare a variety of methods for missing values treatment?
02	Does the paper use more than one public dataset?
03	Does the study use more than one type of assessment measurement?
04	Does the study use statistical tests to assess hypotheses?
05	Does the paper evaluate more than one task of data mining?

Table 2: List of quality criterion used in this systematic review.

Basically, the quality criteria adopted aim to evaluate the studies regarding the methods already available confronting the proposed method with the state of art (criterion 01); the possibility of a reliable comparison between different methods found in the literature (criteria 02 and 03); and analysis robustness (criterion 04).

During the conduction of this systematic review, more than 9,000 publications were found. The large amount of publications available ratify the large capillary of the MV problem, journals from various fields such as psychology, oncology, human behavior and environment, have published studies about methods to handle missing data. Figure 1 shows the sequence of steps applied in this systematic review during the journal selection process.

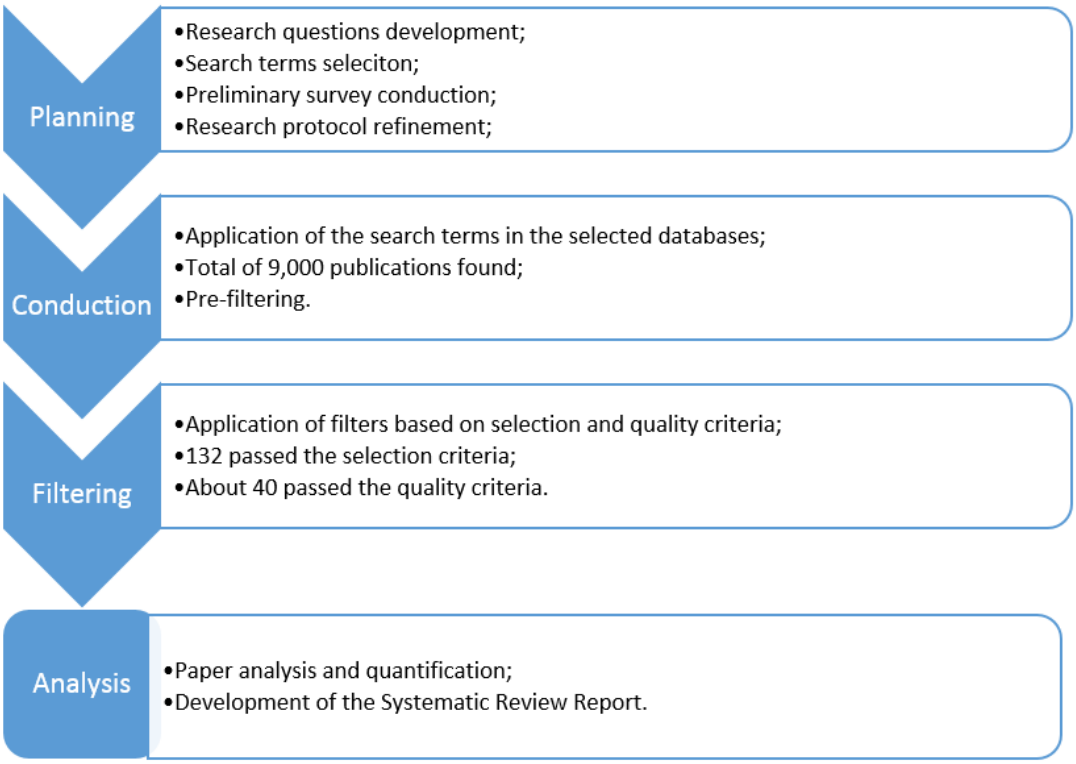


Figure 1: Steps of the process adopted in this Systematic Review.

After filtering the results, 480 studies were analyzed, where 132 papers passed the selection criteria and 40 studies were well ranked based on the quality criteria. Figure 2 shows the systematic review quantification.

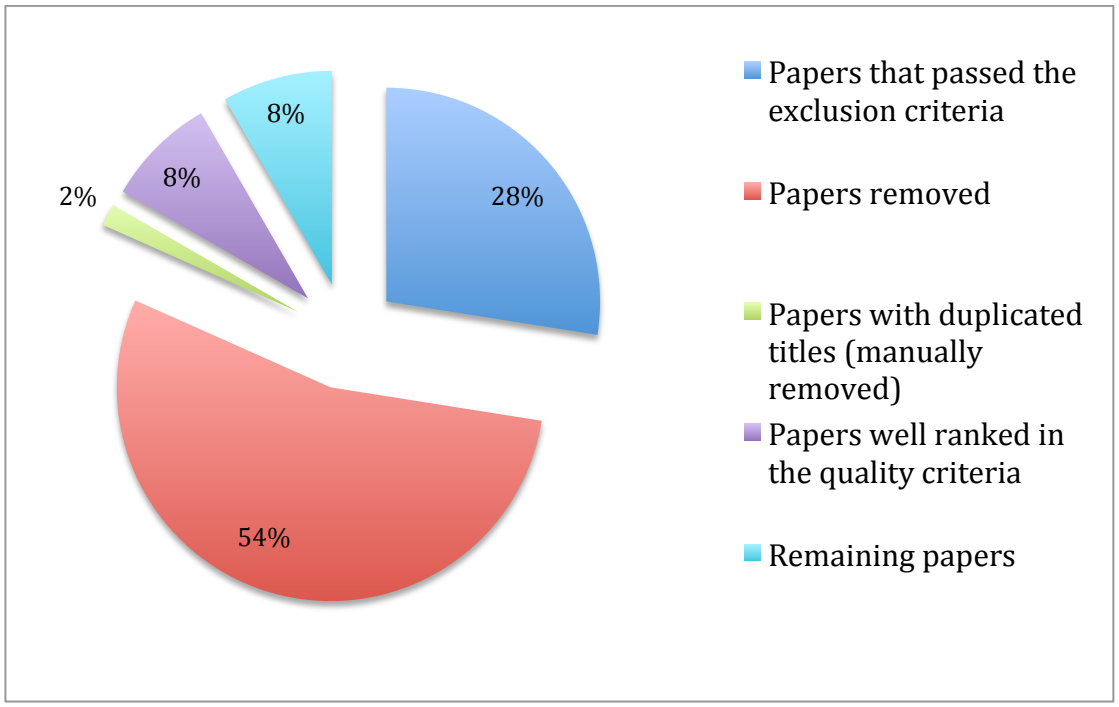


Figure 2: Summary of the systematic review quantifications.

The results of the selected papers analysis are presented in the following section.

4. Discussion

The analysis of the studies that were qualified according to the above criteria was quite diverse, first it was identified trends in dealing with the MV issue, and quantified some of the most important papers published in 2013. 17 papers were analyzed, in which 15 approached the data imputation; 1 provided a literature review; and the other related the robustness of machine learning algorithm regarding missing data. It is important to highlight that not all studies used in this quantification are listed in the 40 papers that were well ranked according to the quality criteria. For example, two studies did not address imputation, failing the quality criteria 04 and 07.

Based on these evidences, it is clear the need to analyze the papers regarding the gaps in the literature, in order to identify researches possibilities. However, despite the fact that it was defined in the research methodology the use of journal papers published only in the last three years, some prior works are also significant.

The review of Schafer and Graham (2002) is a good example; the authors review the state of the art methods indicating points that are still unsolved, giving direction to future researches. The same authors updated the review afterward (Graham, 2009) contemplating an extended theoretical foundation and contextualizing the theory with real problems. The authors also presented a current problem on recent publications, which is the lack of recent treatment methods adoption.

Although the two reviews found were very specific, quantifier tests were performed. For example, the study of García-Laencina et al. (2009) analyzes the MV issue in the pattern classification context; it was made through the presentation and comparison of MVT methods already established in the area, specifically the k-Nearest Neighbor, Multi-Layer Perceptron (MLP), Self-organizing map and expectation-maximization algorithm. By default, the authors define three types of missing data mechanisms, in which the difference lies in the presentation of two possible scenarios, when working with machine learning: a) the dataset used for training is complete while the missing data are only in the test set; b) there are MV in both datasets.

Regarding the performance evaluation, there are different approaches for the classification task in imputation datasets. García-Laencina et al. (2009) consider two types of tasks to be performed: classification and imputation. In the first, since the missing values were imputed, a classifier is trained and its accuracy is measured by the classification error rate. For the imputation, two criteria were used: the predictive accuracy, distance between the actual value and the imputed value; and the distributive accuracy, where the imputation method should preserve the statistical distribution of the real values. The choice of the imputation methods made by García-Laencina et al. (2009) was motivated by the fact that most commercial software for decision support does not handle missing data, which implies the need to treat them externally. The experimental results allow to conclude that there is no universal solution to deliver optimum results in every application domain. Finally, the authors state that in real scenarios, it is necessary to conduct a detailed study to evaluate which estimation method is able to provide an improvement in the classification performance.

The third and final review analyzed addresses MVT strategies in the classification task, specifically about learning algorithms rules for divide-and-conquer strategy (Wohlrab & Fürnkranz, 2010). The proposed method was

tested in 24 UCI datasets that already incorporated MV. For experiment control purposes, three datasets (Credit-G, KRKP and Segment) were arbitrarily chosen, removing values according to the two rules developed by the authors, denominated in the paper as “value amputation”. This procedure ranged according to a removal data rate of 15% in the limits from 15% to 90% for each attribute, not obeying the well-defined mechanisms such as Missing at Random or Missing Completely at Random. The experimental results were evaluated by the average accuracy in relation to the rules generation algorithm.

The reviews listed above are from 2002, 2009 and 2010, in which two reviews by Graham aimed to guide the research on the topic, the others presented the currently most exploited techniques, testing them in controlled experiments. It is known that these reviews are overly restricted whether under the imputation methods comparison or under the dataset diversity. It is reiterated that the literature has no other reviews within two and a half years, which justifies the development of a more comprehensive review considering the missing values mechanisms, as guided by Graham (2009). Additionally, these reviews also lack the utilization of the state of art methods, since the results comparison considered only well-established approaches. It occurs due to the non-adoption of a standard for MVT benchmark experiments, which would increase analyzes diversity.

This prevents experiments replication and affects the studies comparison in the area. In this context, the paper of Luengo, García, and Herrera (2011) is notorious because the authors confront 14 different imputation methods, using 23 classification methods which were divided into three categories: rule induction learning; approximate models; and lazy learning methods. The main evaluation parameter used was the predictive model accuracy.

The main contribution of this study is the correlation of which imputation method is the most applicable to a particular classifiers group, which is accomplished through the Wilcoxon signed-rank hypothesis test (Wilcoxon 1945). Other relevant points referring the analysis influence of the data imputation method related to two measures: Wilson’s noise ratio and average mutual information difference. These measures, despite their usefulness as highlighted in the basic references: only Luengo, Sáez and Herrera (2011), Luengo et al. (2011) and García-Laencina, Sancho-Gómez, and Figueiras-Vidal (2013) used them in fact.

Moving from the review papers to studies focusing on MVT itself, it is noticed a trend that most of the studies developing this line of research intent to contribute to a specific application domain. For example, Chang and Ge (2011) compare imputation techniques applied to data stream; Ding and Ross (2012) compare MVT methods in multibiometric systems by means of imputation; Peyre, Leplège, and Coste (2011) perform a comparison, via simulation, of MVT methods applied to studies on quality of life; Vaden, Gebregziabher, Kuchinsky, and Eckert (2012) adopt the multiple imputation for image signals reconstruction for functional MRI brain scans; Miranda, Krstulovic, Keko, Moreira, and Pereira (2012) and Krstulovic, Miranda, Simoes Costa, and Pereira (2013) use auto associative neural networks (auto encoders) to predict values for imputation in power distribution systems.

Within the studies developing new data imputation methods, Zhang (2010) study is noteworthy, as it provides the following categorization methods based on the imputation numbers:

- Single Imputation: provides a single estimation for each missing value;
- Multiple Imputation: estimates possible values for imputation, based on appropriate measures to verify accuracy, in order to combine them at the final value;
- Fractional Imputation: represents a concession between the first two categories, proposed by Kang, Koehler, and Larsen (2007).
- Iterative Imputation: uses the mechanism of generating and testing considering useful information (including incomplete cases).

On Zhang's (2010) study, the experiments aimed not only at the classification, but also regression. For performance measures, the author reiterated common used measures: the classification accuracy and the RMSE. Regarding the RMSE, the author named the measure as imputation accuracy – corroborating to the lack of standardization in the area.

From a preliminary investigation, it was noted that this measure is consolidated by the assessment of predictive accuracy, as in the works of Little and Rubin (2002); Schafer and Graham (2002); Marwala (2009); Lei and Wan (2010); Dumedah and Coulibaly (2011); Ferrari, Annoni, Barbiero, and

Manzi (2011); Veroneze (2011); Chang, Zhang and Yao (2012); Zhang (2010); Aydilek and Arslan (2013) use the RMSE for this purpose. However, it is important to note that the RMSE refers to the regressors´ accuracy, as in Kang (2013).

Another class of studies worth mentioning, involving MVT, is the one using bio-inspired approaches for data imputation. The analysis concluded that the use of bio-inspired models in the imputation process is still initial, although there is already research in the area. One of the precursors is Abdella and Marwala (2005), when using genetic algorithms to improve the convergence of neural networks; however, it is noticeable that this strategy is old and well established (Montana & Davis, 1989).

From the recent imputation studies involving hybrid approaches with bio-inspired models, the work of Veroneze (2011) and De França, Coelho, and Von Zuben (2013) are worth mentioning; where a technique, based on ant colony and clustering, called bi-clusterization is used. The first study applied the method in a complete base of gene expression, with the missing data being simulated according to all MCAR, MAR e MNAR standards.

The RMSE was also used as measure for the imputation quality assessment; however, it also evaluated the algorithm’s parameters, such as ants quantity in relation to the MV quantity present at the database and the method performance. The authors concluded that the proposed treatment achieved better results compared to KNN Imputation and rSVD (regulated Singular Value Decomposition). It is noteworthy that the methods represent only one classification task, the data clustering.

The study of De França et al. (2013) is presented as an update of the first, where the authors reassured the drawbacks of the first approach, especially regarding how the previous model estimated the MV. Therefore, the work also falls to an analysis of the same measures and methods from the data grouping point of view only.

Some of the studies that approached the MVT by means of imputing data using bio-inspired models are beyond the scope of this discussion, since they did not present a differential approach.

5. Conclusion

The constant increase of computer performance has a great impact in data analysis. In this context, the missing data problem deserves attention due its large high-capillarity. Thus, many treatment methods were developed to reduce the bias imposed by the missing values problem. Because of the large amount of methods available and number of publications, the systematic review process is proven as an essential task to maintain the quality of researches analysis in the area. By applying the systematic review to missing value treatment, it was possible to demonstrate a clear trend in the use of data imputation as a preferred method to handle missing values. In addition, it was noticed a lack in the use of bio-inspired models in the imputation process. Although there are many studies in the area, some modeling flaws were detected and, by these found shortcomings, especially regarding the use of multi-objective approaches, research possibilities are highlighted. Many studies use more than one evaluation measure and some of them conflict with the classification accuracy and the RMSE, which justify the adoption of different multi-objective optimization.

It was also found a lack of standardization in the experiments involving missing values, such as the use of different evaluation measures, the use of bases not available for experiment replication, and non-adoption of statistical tests for hypotheses validation. This gap represents an obstacle on the evaluation, replication and comparison of recently proposed methods, which prevents the full usage of the methods by the academy or industry. Finally, the prospects of this study are to provide important insights for possible researches on the treatment of missing values.

References

Abdella, M. & Marwala, T. (2005). The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database. *Computers and Artificial Intelligence*, 24, 577-589.

Aydilek, I. B. & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233, 25-35.

Brown, M. & Kros, J. (2003). Data Mining. IGI Global, [online] pp.174--198. Available at: http://dl.acm.org/citation.cfm?id=903826.903834.

Chang, G., & Ge, T. (2011). Comparison of missing data imputation methods for traffic flow. *IEEE International Conference on Transportation, Mechanical, and Electrical Engineering* (pp. 639–642).

Chang, G., Zhang, Y., & Yao, D. (2012). Missing Data Imputation for Traffic Flow Based on Improved Local Least Squares. *IEEE Tsinghua Science and Technology, 17(3)*, 304–309.

De França, F. O., Coelho, G. P. & Zuben, F. J. V. (2013). Predicting missing values with biclustering: A coherence-based approach. *Pattern Recognition*, 46, 1255-1266.

Ding, Y. & Ross, A. (2012). A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, 45, 919-933.

Dumedah, G., & Coulibaly, P. (2011). Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *Journal of Hydrology, 400(1-2)*, 95–102.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 37-54.

Ferrari, P. A., Annoni, P., Barbiero, A. & Manzi, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics & Data Analysis*, 55, 2410-2420.

García-Laencina, P. J., Sancho-Gómez, J.-L. & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, 19, 263-282.

García-Laencina, P. J., Sancho-Gómez, J.-L. & Figueiras-Vidal, A. R. (2013). Classifying patterns with missing values using Multi-Task Learning perceptrons. *Expert Systems and Applications*, 40, 1333-1341.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60, 549–76.

Han, J., Kamber, M. (2006). *Data Mining. Concepts and Techniques*. Morgan Kaufmann. ISBN: 1558609016

Heeringa, S., West, B., & Berglund, P. (2010). *Applied survey data analysis*. Chapman & Hall/CRC.

Kang, P. (2013). Locally linear reconstruction based missing value imputation for supervised learning. *Neurocomputing*, 118, 65-78.

Kang, S.-S., Koehler, K., & Larsen, M. D. (2007). Partial FEFI for incomplete tables with covariates. *Iowa State University Press*, Ames.

Kitchenham, B. & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (EBSE 2007-001). Keele University and Durham University Joint Report. (http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8.pdf)

Krstulovic, J., Miranda, V., Simoes Costa, A. & Pereira, J. (2013). Towards an Auto-Associative Topology State Estimator. *IEEE Transactions on Power Systems*, 28(3), 3311-3318.

Lei, K. & Wan, F. (2010). Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau. *IEEE International Conference on Automation and Logistics* (pp. 418–422).

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with missing data.* (2nd ed., p. 408). New York: Wiley.

Luengo, J., Sáez, J. a., & Herrera, F. (2011). Missing data imputation for fuzzy rule-based classification systems. *Soft Computing*, 16(5), 863–881.

Luengo, J., García, S. & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32, 77-108.

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques (1st ed.).* Information Science Reference.

Mcknight, P., Mcknight, K., Sidani, S., & Figueredo, A. (2007). *Missing Data: A Gentle Introduction (Methodology In The Social Sciences).* The Guilford Press.

Miranda, V., Krstulovic, J., Keko, H., Moreira, C. & Pereira, J. (2012). Reconstructing Missing Data in State Estimation With Autoencoders. *IEEE Transactions on Power Systems*, 27(2), 604-611.

Montana, D. J. & Davis, L. (1989). Training Feedforward Neural Networks Using Genetic Algorithms.. In N. S. Sridharan (ed.), *IJCAI* (p./pp. 762-767), : Morgan Kaufmann. ISBN: 1-55860-094-9

Motavalli, J. B. T.-S. (2012). *Mining Big Data: there's gold in those mountains of digital information piling up around us--but how to extract it? The challenges are significant, and the business opportunities huge*. Success. R & L Publishing, Ltd. (dba SUCCESS Media).

Peyre, H., Leplège, A. & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, [online] 20(2), 287-300. Available at: http://dx.doi.org/10.1007/s11136-010-9740-3.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7(2)*, 147–177.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. & Tufano, P. (2012). *Analytics: The Real-World Use of Big Data* (IBM Institute for Business Value - Executive Report). IBM Institute for Business Value. Spolaôr, N., Cherman, E., Metz, J., & Monard, C. (2013). A systematic review on experimental multi-label learning. *ICMC Database of Systematic Reviews, 2013*(392), 1-31. http://labic.icmc.usp.br/?q=node/830.

Vaden, K. I., Gebregziabher, M., Kuchinsky, S. E., & Eckert, M. a. (2012). Multiple imputation of missing fMRI data in whole brain analysis. *NeuroImage, 60(3)*, 1843–55.

Veroneze, R., de França, F. O. & Zuben, F. J. V. (2011). Assessing the performance of a swarm-based biclustering technique for data imputation. *IEEE Congress on Evolutionary Computation*, (p./pp. 386-393), : IEEE.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1(6)*, 80–83.

Wohlrab, L. & Fürnkranz, J. (2011). A review and comparison of strategies for handling missing values in separate-and-conquer rule learning. *Journal of Intelligent Information Systems*, 36, 73-98.

Zhang, S. (2010). Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, 35(1), 123–133.

Ackowlegdements

The authors would like to thank CNPq and grant PROCAD-NF/CAPES and PROPESP/UFPA for the support to this research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.