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

TESE DE DOUTORADO

Cláudio Alex Jorge da Rocha

Estratégia de Otimização para a Melhoria da Interpretabilidade de Redes Bayesianas: Aplicações em Sistemas Elétricos de Potência

UFPA/ITEC/PPGEE
BELÉM - 2009

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

TESE DE DOUTORADO

Cláudio Alex Jorge da Rocha

Estratégia de Otimização para a Melhoria da Interpretabilidade de Redes Bayesianas: Aplicações em Sistemas Elétricos de Potência

Tese submetida à Banca Examinadora do Programa de Pós-graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Doutor em Engenharia Elétrica, elaborada sob a orientação do Prof. Dr. Carlos Renato Lisboa Francês.

UFPA/ITEC/PPGEE
BELÉM - 2009

Estratégia de Otimização para a Melhoria da Interpretabilidade de Redes Bayesianas: Aplicações em Sistemas Elétricos de Potência

Esta tese foi julgada adequada para o Exame de Defesa de Doutorado em Engenharia Elétrica, na área de Computação Aplicada, e aprovado na sua forma final pela banca examinadora designada pelo Programa de Pós-Graduação em Engenharia Elétrica do Instituto de Tecnologia da Universidade Federal do Pará em 10 de dezembro de 2009.

.....
Prof. Dr. Carlos Renato Lisboa Francês(UFPA) - Orientador

.....
Prof. Dr. Ubiratan Holanda Bezerra (UFPA) - Membro

.....
Prof. Dr. Maria Emília de Lima Tostes (UFPA) – Membro

.....
Profa. Dra. Solange Oliveira Rezende(USP-São Carlos) – Membro externo

.....
Prof. Dr. Nandamudi.Lankalapali Vijaykumar (LAC-INPE) – Membro externo

VISTO:

.....
Prof. Dr. Marcus Vinícius Alves Nunus
Coordenador do UFPA/ITEC/PPGEE

UFPA / ITEC / PPGEE
BELÉM/PA
2009

R672e Rocha, Cláudio Alex Jorge da

Estratégia de otimização para a melhoria da interpretabilidade de redes bayesianas: aplicações em sistemas elétricos de potência / Cláudio Alex Jorge da Rocha; orientador, Carlos Renato Lisboa Francês - 2009.

Tese (Doutorado) – Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2009.

1. Exploração de dados (computação). 2. Teoria bayesiana de decisão estatística. 3. Energia elétrica – consumo. I. Orientador. II Título.

CDD 22. ed. 005.74

Dedicatória

À **Cleise** e **Sofia**, motivação e razão maior de todos os meus passos. A vocês, minhas Gatinhas, devo e dedico tudo.

Agradecimentos

À Cleise e Sofia, minhas inspirações superlativas, pela compreensão, incentivo e todo amor em cada momento da minha vida. O desenho dos meus sonhos e da realidade de minhas ações é efetivado com os traços e as cores da cumplicidade de nossos desejos e realizações.

Aos meus pais, Chico e Gláucia, esteio da minha formação, principalmente pela simplicidade e alegria com que realizam suas vidas. Como se não bastasse agradecer pela própria existência, deixo registrada a minha eterna gratidão pelo amor incondicional que sentem por nós.

Aos meus irmãos, Alexandre, Alexsandro, Ana Cláudia, Eduardo e Ricardo, pelo amor fraternal alicerçado no respeito às nossas diferenças, opiniões, trajetórias de vidas, razão maior da nossa união e da imperativa presença, juntos, na órbita de nossos pais.

A todos que fazem a nossa grande família, sobrinhos, tios, primos, sogro, sogra, cunhados. Pela convivência e importância que cada um tem nesse universo social e, em consequência, das experiências, orientações, carinho e afeto compartilhados.

Ao Renato, amigo de todas as jornadas. Vou usar um termo teu "É impressionante" como, ao longo desses mais de 27 anos de amizade, sempre tu estivestes presente nas principais lembranças, conquistas e histórias. Da bola no CEPC às festas do Círculo Militar, das paródias à fundação do I.R.A, do vestibular à realização da graduação, do Mestrado ao Doutorado. "É disso que estou falando", do fortalecimento e da longevidade da nossa amizade, da cumplicidade nas realizações dos nossos sonhos, ainda dos tempos de moleque, em que os nossos caminhos poderiam ter sido bem diferentes.

Ao Ádamo, por suas imensas contribuições, comprometimento e sapiência no apoio às escolhas dos melhores caminhos para a realização desta tese.

Ao Cláudinho, Diego, Edvar, Jorge, Liviane e Marcelino pela convivência sempre prazerosa no LPRAD, PRODEPA e em todos os cantos da vida em que nossa amizade se solidifica.

Ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará, em especial seus professores, pelo ambiente profícuo e fértil ao progresso de minha formação acadêmica.

À Universidade da Amazônia, que sempre incentivou e apoiou a minha trajetória acadêmica.

Ao Afonso, André, Dário, Fabrício, Paulo e Ricardo – o sexteto bicolor. Pelo companheirismo, respeito e, claro, as geladas que sempre regaram nossa amizade.

Ao Prof. Eloi, por suas contribuições e compreensão pelos percursos que foram tomados para a consecução desse trabalho.

À Solange, além de nossa amizade, por construir o pilar principal da formação na área que hoje milito. Se gosto e realizo algo nessa área é por conta de seus ensinamentos e orientações.

A Deus, pela vida.

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AG	Algoritmo Genético
ANEEL	Agência Nacional de Energia Elétrica do Brasil
CCEE	Câmara de Comercialização de Energia Elétrica
CELPA	Centrais Elétricas do Estado do Pará
DGA	*Dissolved Gas Analysis*
DPC	Distribuição de Probabilidade Conjunta
KDD	*Knowledge Discovery in Database*
LPRAD	Laboratório de Planejamento de Redes de Alto Desempenho
MAP	*Maximum a Posteriori*
MCSD	Mecanismo de Compensação de Sobras e Déficits de Energia
MD	Mineração de Dados
MDL	Descrição de Comprimento Mínimo
MRSL	*Multiple Regression Structure Learner*
PLD	Preço de Liquidação das Diferenças
RB	Redes Bayesianas
TPC	Tabelas de Probabilidades Condicionais

Resumo

A investigação de métodos, técnicas e ferramentas que possam apoiar os processos decisórios em sistemas elétricos de potência, em seus vários setores, é um tema que tem despertado grande interesse. Esse suporte à decisão pode ser efetivado mediante o emprego de vários tipos de técnicas, com destaque para aquelas baseadas em inteligência computacional, face à grande aderência das mesmas a domínios com incerteza. Nesta tese, são utilizadas as redes Bayesianas para a extração de modelos de conhecimento a partir dos dados oriundos de sistemas elétricos de potência. Além disso, em virtude das demandas destes sistemas e de algumas limitações impostas às inferências em redes bayesianas, é desenvolvido um método original, utilizando algoritmos genéticos, capaz de estender o poder de compreensibilidade dos padrões descobertos por essas redes, por meio de um conjunto de procedimentos de inferência em redes bayesianas para a descoberta de cenários que propiciem a obtenção de um valor meta, considerando a incorporação do conhecimento *a priori* do especialista, a identificação das variáveis mais influentes para obtenção desses cenários e a busca de cenários ótimos que estabeleçam valores, definidos e ponderados pelo usuário/especialista, para mais de uma variável meta.

Palavras-chave: *Mineração de Dados; Redes Bayesianas; Inferência Bayesiana; Algoritmos Genéticos; Descoberta de Cenários; Sistemas Elétricos de Potência.*

Abstract

The study of methods, techniques and tools that can aid the decision processes in power systems, in its many sections, is a subject of great interest. This decision support can be accomplished through many different techniques, particularly those based on computational intelligence, given their applicability on domains with uncertainty. In this proposal, Bayesian networks are used for the extraction of knowledge models from the available data on power systems. Moreover, given the demands of these systems and some limitations imposed to the inferences in Bayesian networks, a method is proposed, using genetic algorithms, capable of extending the power of comprehensibility of the patterns discovered; it aims at finding the optimal scenario in order to attain a given target, considering the incorporation of a priori knowledge from domain specialists, identifying the most influent variables in the domain for the maximization of the target variable.

Keywords*: Data Mining; Bayesian networks; Bayesian Network; Bayesian Inference; Genetic Algorithm; Scenario Discovery; Power Systems.*

1. INTRODUÇÃO

Ao longo da história, o setor elétrico tem desempenhado papel importante no desenvolvimento de qualquer região ou país. Em países em desenvolvimento como o Brasil, e mais ainda, em regiões periféricas, em termos econômicos, como a Amazônia, o provimento estável e de qualidade de energia elétrica é condição *sine qua non* para o progresso e a conseqüente inclusão social da população que habita essa Região.

Os desafios das empresas de geração e distribuição de energia, pela eficiência energética, no que tange aos processos, equipamentos e ações/métodos, têm crescido consideravelmente no mundo devido a vários motivos, dentre eles, auferirem vantagens econômicas, evitarem a degradação das condições ambientais locais e regionais, reduzirem as mudanças climáticas associadas ao efeito estufa, promover o desenvolvimento sustentável e criar uma cultura junto à sociedade de combate ao desperdício de energia elétrica.

Com isso, a investigação de métodos, técnicas e ferramentas que possam apoiar os processos decisórios dos sistemas elétricos, em seus vários setores, é um tema que tem despertado grande interesse à pesquisa nacional e internacional. Esse suporte à decisão pode ser efetivado mediante o emprego de vários tipos de métodos, entretanto os sistemas inteligentes emergem como um dos que apresentam resultados mais robustos.

Além disso, nos dias de hoje, em que o mercado deste e dos demais setores da economia se apresentam cada vez mais competitivo, as organizações têm de modo imperioso recorrido às suas bases de dados e as informações externas que a influenciam, para extrair padrões que melhor expressem suas ações futuras, como forma de aperfeiçoar os seus processos decisórios e, em decorrência, o andamento de seu negócio.

A preocupação cada vez maior das empresas em investirem em tecnologias de armazenamento e processamento de informações, bem como de transformá-las em

conhecimento, tem atraído interesse pela investigação de mecanismos que possam realizar essa transformação, proporcionando um auxílio efetivamente inteligente ao processo de tomada de decisão.

Essas investigações podem ser enquadradas em uma área denominada Mineração de Dados (MD), também conhecida como Extração de Conhecimento de Base de Dados (KDD – *Knowledge Discovery in Database*), que representa uma fonte de maduras tecnologias, amplamente incorporadas nos processos organizacionais dessas corporações. A MD pode ser compreendida como um processo interativo e iterativo para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (Fayyad et al., 1996).

O sucesso da aplicação do processo de MD está intimamente ligado a diversos aspectos, desde o entendimento do domínio da aplicação, passando pela qualidade dos dados a partir dos quais a extração de conhecimento vai ser realizada e das escolhas das técnicas e ferramentas utilizadas, até a facilidade de compreensão dos padrões descobertos. Portanto, é pouco provável que os padrões descobertos por meio de dados inconsistentes ou com valores ausentes, sem que haja um devido tratamento desses problemas, garantam aos gestores decisões confiáveis. Assim como a confiabilidade e compreensibilidade dos padrões descobertos, decorrem de escolhas bem feitas das tarefas e técnicas de mineração utilizadas.

Este trabalho está pautado na investigação da melhoria da compreensibilidade dos padrões descobertos no processo de MD. Segundo Rezende (2003), a compreensibilidade desses padrões diz respeito à facilidade de interpretação dos mesmos por um ser humano. Assim, a utilização de técnicas de MD para proverem mecanismos de apresentação e visualização, que simplifiquem a análise do conhecimento obtido, pode contribuir fortemente para que os usuários possam medir a qualidade deste conhecimento. A qualidade é freqüentemente aferida pelo grau de interesse que o modelo obtido pode ter para o usuário (Chen, 2001).

Dentre as diversas técnicas de MD encontradas na literatura, é possível destacar uma das mais proeminentes, quando se trata de facilidade de interpretação do conhecimento obtido em um domínio com incerteza – as redes Bayesianas (RB), visto que essa técnica

provê um mecanismo de representação do modelo causal de um determinado conjunto de dados (Pearl, 1988), permitindo análises qualitativas e quantitativas entre as varáveis deste domínio e, desse modo, dando suporte ao processo de tomada de decisão a partir dos seguintes tipos de inferências: diagnóstico, causais, intercausal e da combinação entre esses tipos (Russel e Norvig, 2003); (Korb e Nicholson, 2003).

Entretanto, as redes Bayesianas apresentam limitações quando se pretende estabelecer qual a combinação ótima dos estados de determinadas variáveis (discretas ou contínuas) que geram um determinado requisito alvo (estado de outra variável do domínio). Em muitas aplicações reais, a busca por situações que expliquem a obtenção de determinadas metas é recorrente. Por exemplo, para se alcançar um determinado índice de vendas, é necessário encontrar qual o conjunto de fatores que podem influenciá-las e, assim, determinar quais os valores (estados) das variáveis que compõem esses fatores exercem maior impacto no índice de vendas a ser obtido. A motivação maior para a consecução deste trabalho é suprir essa restrição das RBs, o que é feito a partir da combinação das técnicas de algoritmo genético com as RBs, obtidas a partir de bases dados.

É possível delinear como objetivo basilar deste trabalho a elaboração de uma estratégia capaz de estender o poder de interpretabilidade das RBs, geradas para medir o relacionamento causal entre as variáveis de um determinado domínio, por meio da descoberta de cenários, isto é, encontrar os valores que compõem uma combinação (configuração) ótima dos estados, de um conjunto de variáveis desse domínio, que possibilitem a obtenção de um valor meta de uma determinada variável que não compõe o referido conjunto.

Para efeito de estudo de caso, o método aqui desenvolvido é aplicado no setor elétrico, mais especificamente para encontrar os cenários sócio-econômicos e climáticos que contribuem para a obtenção de um valor meta de consumo de energia, em suas várias classes (residencial, comercial, etc.).

A descoberta desses cenários é de extrema valia para o setor elétrico, posto que propiciam a análise das correlações entre os fatores sócio-econômicos e climáticos que influenciam o consumo de energia elétrica, com vistas ao planejamento e operação desse setor de maneira confiável e segura.

Além disso, essas análises servem de suporte à tomada de decisões, por parte das concessionárias de energia elétrica, relacionadas ao mercado de energia, em razão de possibilitarem a antecipação das condições favoráveis à obtenção de determinadas metas de consumo e, em consequência, permitem elaborar processos de comercialização de energia elétrica mais adequados, visto que previsões aquém do necessário, deixam as concessionárias sujeitas a multas e ainda obrigadas a celebrar contratos de energia de curto prazo, normalmente mais custosos, para atender a demanda dos consumidores. Do mesmo modo, as previsões além do necessário, também são passíveis de penalidades pela Agência Nacional de Energia Elétrica (ANEEL).

À luz desses indicativos, podem-se destacar como principais objetivos específicos desta tese, os que seguem:

- Aplicar o processo de MD em um domínio real, mais especificamente no mercado de energia elétrica. Como já mencionado, a idéia da aplicação de RBs neste setor é propor um sistema de suporte à decisão que permita a tomada de ações de maneira antecipada, baseada na exploração dos cenários que o método desenvolvido pode projetar;
- Empregar técnicas híbridas de Inteligência Computacional com vistas à melhoria dos processos de descoberta de conhecimento em bases de dados;
- Criar um método o mais genérico possível no que concerne à complexidade da RB obtida a partir dos dados de um determinado domínio de aplicação. Em virtude de o método híbrido desenvolvido utilizar o mecanismo de inferência da RB, em seu processo de descoberta de cenários, é possível o emprego tanto de métodos exatos quanto aproximados de inferência, sendo o último fundamental quando se está trabalhando com RBs de alta complexidade, cujas inferências são algumas vezes intratáveis por meio de métodos exatos;
- Permitir a análise de cenários antecipatórios em RBs, ao contrário de como essas redes são frequentemente utilizadas para a obtenção de cenários exploratórios. O método desenvolvido nesta tese procura traçar diretrizes para que determinada situação seja alcançada ou evitada. Neste tipo de estudo, se definem metas a serem obtidas e, assim, os cenários representam que tipo de estratégia, decisões

ou configuração dos valores das variáveis do sistema analisado podem levar até a meta alvo, definida no início do processo;

- Desenvolver soluções capazes de identificar as variáveis de um domínio que mais contribuem para obtenção de uma meta estabelecida;
- Obter, para efeito da descoberta de cenários, métodos que levem em consideração o conhecimento *a priori* do especialista do domínio;
- Construir soluções para a descoberta de cenários que considerem mais de uma meta estabelecidas pelo usuário.
- Adequar ainda mais os sistemas de Suporte à Decisão baseados em MD a aplicações do mundo real, provendo esses sistemas de outros modos de interpretação e inferências.

A partir dos objetivos deste trabalho, algumas hipóteses podem ser suscitadas. A principal delas é que a estratégia desenvolvida seja capaz de estender eficientemente o poder de interpretabilidade das RBs. Além dessa, outras hipóteses podem ser levantadas, o que é descrito abaixo:

- A utilização de técnicas de otimização permitem aperfeiçoar o processo de inferência das RBs;
- O emprego de técnicas híbridas de Inteligência Computacional propicia a melhoria dos processos de descoberta de conhecimento em bases de dados;
- É possível utilizar RBs para a descoberta de cenários antecipatórios;
- Empregando-se técnicas híbridas de inteligência computacional é viável a identificação de cenários com mais de um valor meta;
- O emprego de técnicas híbridas de inteligência computacional permite identificar quais variáveis exercem maior influência na consecução de um determinado cenário;
- A possibilidade de incorporação de conhecimento *a priori* do especialista do domínio melhora os processos de descoberta de conhecimento em bases de dados;

As hipóteses levantadas podem ser corroboradas pela estratégia criada nesta tese para melhoria da interpretabilidade das RBs, fundamentalmente em razão das seguintes contribuições, que foram concebidas a partir das investigações que alicerçaram o desenvolvimento da abordagem decorrente dos estudos realizados para a realização desta Tese.

- O desenvolvimento de uma estratégia para descoberta de cenários, por meio da combinação de técnicas bayesianas e evolucionárias, que possibilitem a obtenção de um valor meta de uma determinada variável que compõe o domínio de aplicação.
- facilidade de uso de métodos de inferência em redes bayesianas exatos e aproximados como função de avaliação dos cenários;
- possibilidade de encontrar as variáveis que exercem maior influência na obtenção de um valor meta;
- capacidade de se obter os valores contínuos das variáveis de evidência;
- incorporação de conhecimento do especialista para conduzir o processo de busca da melhor configuração das variáveis de consulta;
- descoberta de cenários (configuração mais provável) com mais de uma variável meta, ponderadas pelo especialista, no contexto do domínio de aplicação.
- Ao contrário do modo como as técnicas evolucionárias são utilizadas na maioria dos sistemas que envolvem RBs propostos na literatura, em que as mesmas são adotadas para otimizar o processo de aprendizado das estruturas dessas redes, neste trabalho, os conceitos de algoritmos genéticos são empregados para estender o poder de inferência das RBs, melhorando a compreensibilidade dos padrões obtidos por estas redes.

Para um melhor entendimento dos temas aqui tratados, este trabalho está organizado conforme é apresentado a seguir.

No Capítulo 2, são apresentados os conceitos relacionados aos sistemas de suporte à decisão baseados em processos de MD, como as motivações para o emprego nesse tipo de

sistema, as etapas que compõem os processos de MD, bem como as principais fronteiras de pesquisas nessa área.

No Capítulo 3, são descritas as técnicas utilizadas em MD para manipulação de incerteza, com ênfase nos conceitos associados às RBs.

No Capítulo 4, são focalizados os trabalhos correlatos, a partir dos quais é possível estabelecer paralelos com a tese ora realizada.

No Capítulo 5, é apresentado o método de descoberta de cenários, baseado na estratégia de otimização para a melhoria da interpretabilidade de redes bayesianas.

No Capítulo 6, é focalizada a aplicação do método no contexto dos sistemas elétricos de potência.

Finalmente, no Capítulo 7, são elencadas as conclusões elaboradas a partir desta tese, bem como os trabalhos futuros decorrentes dos estudos aqui realizados.

2. SISTEMAS DE SUPORTE À DECISÃO BASEADOS EM MINERAÇÃO DE DADOS

2.1. CONSIDERAÇÕES INICIAIS

O Suporte à decisão está associado freqüentemente a aplicações envolvendo análise e exploração dos dados e informações históricas de uma organização, de modo a prover um mecanismo de alto nível para auxiliar os processos de tomada de decisão. Segundo Chen (1999) e Turban (2001), esses mecanismos podem ser utilizados, basicamente, nas seguintes situações:

- Alerta a usuários;
- Reconhecimento de problemas;
- Solução de problemas;
- Planejamento de estratégias;
- Aquisição, transformação e exploração de novos conhecimentos;
- Simulação;
- Facilitar a interação entre os usuários de níveis decisórios.

É imperativo que esses mecanismos sejam implementados em sistemas computacionais face à grande quantidade de dados que são manipulados e à própria condição humana, relacionada com a limitação para a análise de grande volume de informações, estresse, erro, desatenção, etc. As soluções computacionais utilizadas são compostas, em sua maioria, por um conjunto de métodos, técnicas e ferramentas de áreas como bancos de dados, estatística, e inteligência computacional. Um típico e proeminente exemplo dessas soluções se enquadra na área denominada Mineração de Dados.

Neste capítulo, são apresentados os conceitos básicos do processo de MD sob a ótica de que esse processo é o núcleo dos sistemas inteligentes de apoio à tomada de decisão.

2.2. MINERAÇÃO DE DADOS

A MD é um termo usado para descrever o processo de descoberta de conhecimento em bases de dados e consiste de uma tecnologia que combina métodos e ferramentas de estatística, inteligência computacional e banco de dados para encontrar uma descrição matemática e/ou lógica, eventualmente complexa, de padrões e regularidades nos dados (Fayyad, 1996).

Apesar de alguns autores considerarem a MD como uma etapa do processo KDD, neste trabalho, os termos MD e KDD serão tratados indistintamente.

Segundo Chen (2001), o entendimento da aplicação de um processo de MD pode ser obtido a partir de oito elementos ou primitivas: P (especificação do problema), D (amostra representativa dos dados), T (a tarefa de MD utilizada), F (Conhecimento de fundo), M (algoritmo ou técnica de MD), C (modelo ou padrão extraído, juntamente com sua forma de representação e visualização), A (Avaliação do Modelo, permitindo a mensuração da simplicidade, precisão e utilidade do conhecimento extraído) e U (usuários). Assim, o processo de MD pode ser descrito pela 8-tupla: (P, D, T, F, M, C, A, U). O relacionamento entre os elementos pode ser compreendido como se segue. O usuário (U) define o problema (P), em seguida a MD é realizada sobre os dados selecionados na amostra representativa da base de dados (D) e o conhecimento a priori do problema (F). Essa mineração é realizada a partir da definição da tarefa de mineração (T) utilizada (e.g. classificação ou associação) e do algoritmo de MD (M), que representam o núcleo do processo de extração de conhecimento. Por fim, esses padrões (C) obtidos são avaliados (A) e retornados para o usuário.

Desta forma, a partir dessas primitivas, a descoberta de conhecimento a partir de dados é entendida como um processo contendo, pelo menos, as seguintes grandes etapas: (1) Compreensão do Domínio da Aplicação; (2) pré-processamento dos dados; (3) Extração

de Padrões (4) Avaliação do Conhecimento Extraído ou Pós-Processamento e (5) Consolidação e Utilização do Conhecimento Extraído. Um esquema representativo contendo todos essas etapas é ilustrado na figura 2.1.

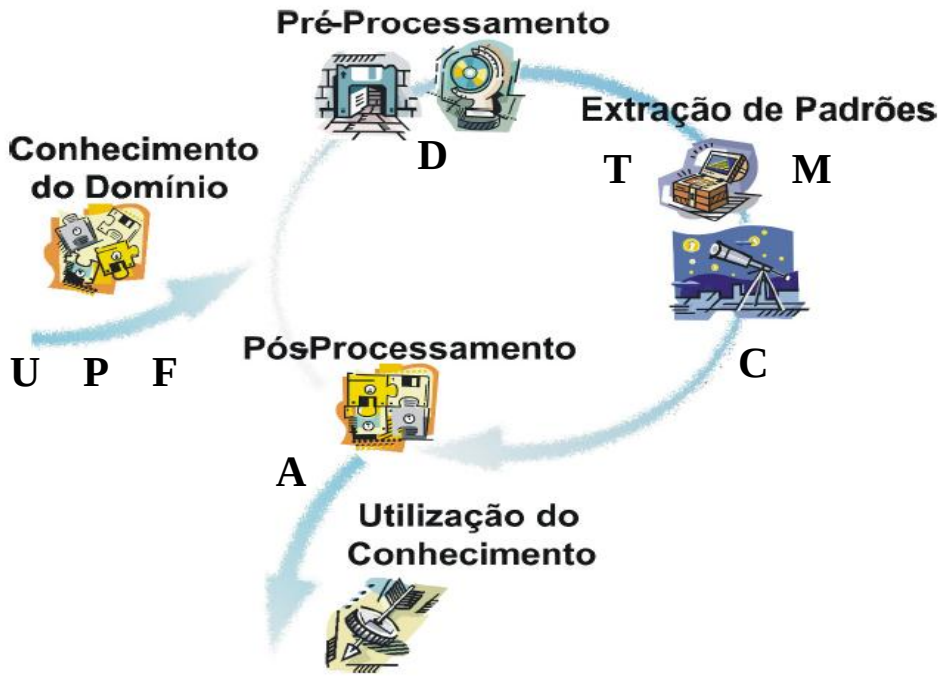

Figura 2.1. Etapas do processo de mineração de dados. Adaptado de (Rezende, 2003).

O processo de MD inicia com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados. Em seguida, uma amostra representativa é retirada (e.g. utilizando técnicas estatísticas) da base de dados, pré-processada e submetida aos métodos e ferramentas da etapa de Extração de Padrões com o objetivo de encontrar padrões/modelos (conhecimento) a partir dos dados. Posteriormente, esse conhecimento é avaliado quanto a sua qualidade e/ou utilidade para que, em caso positivo, seja utilizado para apoio a algum processo de tomada de decisão.

É importante notar que, por ser um processo eminentemente iterativo, as etapas do processo de MD não são estanques, ou seja, a correlação entre as técnicas e métodos utilizados nas várias etapas é considerável, a ponto da ocorrência de pequenas mudanças em uma delas afetar substancialmente o sucesso de todo o processo. Desta forma, os resultados obtidos em uma determinada etapa podem acarretar mudanças a quaisquer das

etapas anteriores ou ainda, o recomeço de todo o processo (Rocha, 1999). Um detalhamento maior das metas de cada etapa será apresentado a seguir.

2.2.1. Compreensão do domínio da aplicação

Nesta etapa, busca-se, primordialmente, o levantamento dos requisitos do domínio e a definição das metas a serem alcançadas com a aplicação do processo de MD, considerando aspectos como os apresentados em Rezende (2003):

- Estudo de viabilidade e custos da aplicação do processo;
- Verificação da quantidade e do tipo de conhecimento disponível antes do processo iniciar;
- Análise das condições e metas do usuário final;
- Análise da relação entre simplicidade e precisão do conhecimento a ser extraído.

Além desse levantamento inicial, a compreensão do domínio impacta no sucesso de todas as demais etapas, posto que, a partir desse entendimento e das interações com o especialista do domínio ao longo do processo de MD, é possível subsidiar uma série de ações vindouras, tais como a seleção de características (atributos), ajuste de parâmetros dos algoritmos de MD, bem como prover ao analista uma capacidade maior para efetuar juízo de valor dos padrões descobertos.

2.2.2. Pré-processamento dos dados

A extração direta de padrões a partir de grandes volumes de dados pode se tornar uma tarefa inviável. Grandes volumes de dados podem gerar um espaço de busca de padrões combinatorialmente explosivo. Em adição, as limitações de tempo de processamento e de memória são fatores impeditivos para a submissão direta dos dados, sem o estabelecimento de qualquer filtro, aos algoritmos de MD (Rezende, 2003).

A busca de conhecimento em grandes bases de dados pode ocasionar, ainda, o aumento das chances de se encontrarem padrões pouco significativos e até mesmo espúrios. Uma possível solução para esses problemas envolve a tentativa de selecionar uma amostra

significativa da base de dados da aplicação, além de prover mecanismos de limpeza e transformação dos dados, com vistas a melhorar a qualidade dos mesmos, bem como adequá-los aos objetivos propostos para a aplicação de MD ou às limitações e requisitos das técnicas de MD utilizadas (Tan et al., 2005).

Um conjunto de fatores corrobora para a realização deste pré-processamento, são eles:

- Os bancos de dados atuais costumam ser muito grande e submetê-los a uma ferramenta de MD não seria viável, pois muitas dessas ferramentas possuem limitações e especificidades quanto à entrada de dados. Além disso, por mais que a ferramenta suportasse um grande volume de dados, a complexidade do conhecimento adquirido propiciaria uma difícil compreensão, bem como a possibilidade da obtenção de padrões espúrios comprometeria sobremaneira a utilidade do conhecimento obtido;
- A ausência de valores de atributos dos registros de uma base de dados;
- A presença de ruídos nos dados, caracterizado por valores incorretos que são inseridos no banco (e.g. por falha humana) durante a entrada dos dados, erro de transmissão de dados, entre outros;
- A inconsistência dos dados, resultado do uso de diferentes convenções ou nomes para a representação de um dado;
- A necessidade de transformações nos dados (e.g. normalização) para atender as necessidades das ferramentas de Mineração dos Dados e/ou de análise dos mesmos.

Os aspectos supracitados podem ser tratados por diversas técnicas de pré-processamento de dados, que, em grande parte da literatura, são classificadas em :

- Limpeza dos dados, com objetivo de tratar os valores ausentes e ruídos nos dados, além de corrigir problemas relativos à inconsistência;
- Integração e transformação dos dados, utilizada quando os dados que serão submetidos à ferramenta de MD são provenientes de múltiplas fontes, evitando

redundância de dados, bem como a criação de novos atributos a partir de outros, normalização de valores, entre outras tarefas;

- Redução dos Dados, eliminando atributos que não são relevantes para a tarefa de mineração de dados, além de prover mecanismos de discretização dos dados.

Na figura 2.2, é apresentado um esquema representativo das principais tarefas mencionadas para o pré-processamento de dados.

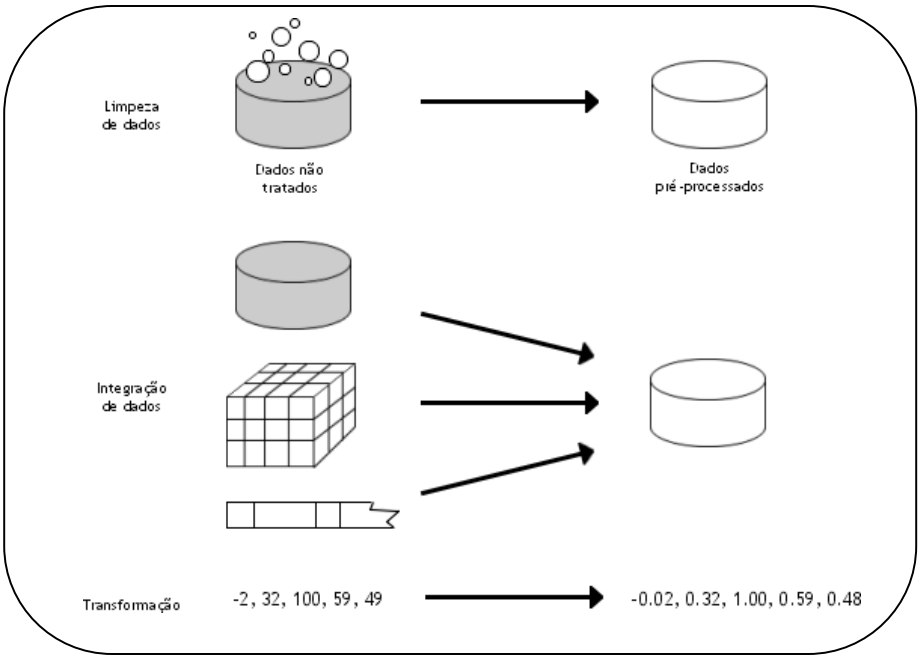

Figura 2.2. Técnicas que podem ser utilizadas no pré-processamento dos dados (Han e Kamber, 2001).

Han e Kamber (2001) salientam, ainda, que apesar de algumas ferramentas de mineração de dados possuírem rotinas para pré-processamento que tentam fazer este tratamento, estas rotinas nem sempre são eficientes, pois não fazem parte do escopo principal da ferramenta.

Uma vez pré-processado, o conjunto de dados é submetido a um (ou mais) algoritmo de MD, onde padrões, comportamentos e regularidades nos dados são encontrados. As ferramentas geram então, como saída, padrões que podem ser interpretados pelo usuário especialista do domínio ou de forma automática por um computador.

2.2.3. Extração de Padrões

Essa etapa do processo de MD corresponde basilarmente à escolha dos modelos empregados para a realização da extração de padrões propriamente dita. Essa escolha está pautada, dentre outros fatores, à linguagem de representação utilizada e no tipo de tarefa de MD eleita. A linguagem para representar os conceitos (padrões) geralmente determina a flexibilidade do modelo em representar conhecimento e a facilidade de compreensão do modelo em termos humanos. Na literatura, podem-se encontrar vários tipos de representação, entre as quais se destacam: árvores e regras de decisão, modelos baseados em casos e modelos de dependência gráfica probabilísticas (redes bayesianas), empregados nesta tese.

As tarefas básicas, realizadas via processo de MD, são comumente classificadas em duas categorias:

- Descritivas: concentram-se em encontrar padrões que descrevam os dados, caracterizando as propriedades gerais desses dados, de forma interpretável pelos seres humanos.
- Preditivas: realizam inferência nos dados correntes para construir modelos, que serão utilizados para reedições do comportamento de novos dados.

Na figura 2.3, são apresentadas as principais tarefas de MD e suas classificações quanto às categorias supracitadas.

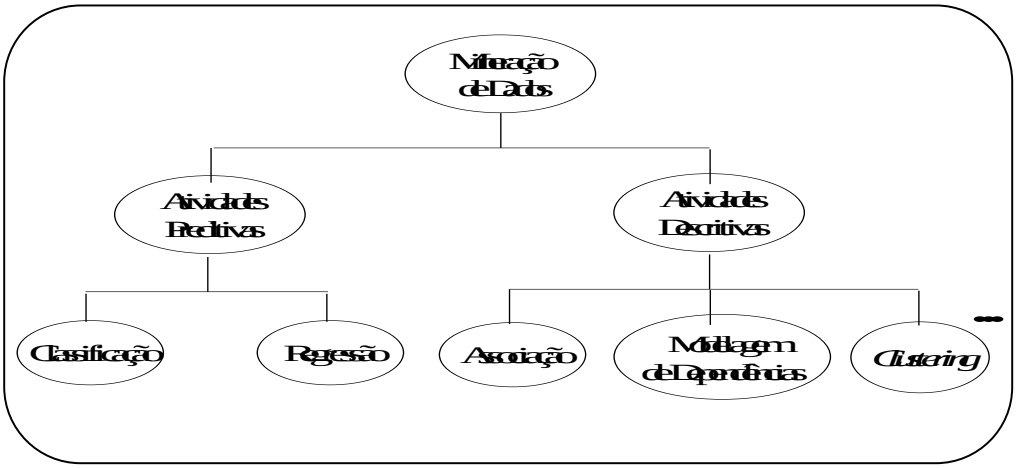

Figura 2.3. Principais tarefas de mineração de dados (Rezende 2003).

A classificação visa o mapeamento de um determinado caso (registro do conjunto de dados) dentro de uma das várias classes pré-definidas. (e.g. regras de classificação a respeito de doenças podem ser extraídas de um conjunto de casos conhecidos e usadas para fazer diagnóstico em novos pacientes, baseados em seus sintomas). De modo simplificado, a tarefa de regressão é conceitualmente a mesma da classificação, sendo que na atividade de classificação, a classe é categórica, enquanto na regressão, a classe representa um valor contínuo (e.g. determinar o consumo de energia elétrica de uma determinada unidade consumidora em um mês futuro).

As regras de associação determinam as relações entre os atributos de uma base de dados (e.g. regras de associação podem descrever que itens são comumente comprados juntamente com outros em um supermercado).

Outra tarefa descritiva a ser destacada é a de *clustering*. Também conhecida como análise de agrupamentos, essa tarefa permite criar, a partir de grandes conjuntos de dados, agrupamentos de dados menores e com características semelhantes. Nas tarefas de *clustering*, não há necessidade da determinação prévia de classes (e.g. investigar quais os diferentes grupos de consumidores de drogas, com base em suas características pessoais).

Por fim, a tarefa de modelagem de dependência, na qual estão incluídas as RBs, permite codificar o relacionamento probabilístico entre os atributos de um conjunto de dados. O uso de modelos dessa categoria de tarefa é o foco deste trabalho.

A partir da definição da tarefa de MD a ser realizada, é possível determinar qual tipo algoritmo será utilizado para realizar o processo de extração propriamente dito. Os modelos induzidos dos dados seguem geralmente os padrões estatísticos, neurais, simbólicos, de dependências probabilísticas ou baseados na teoria de algoritmos genéticos.

Um modelo estatístico típico é gerado pelo método de regressão (e.g. regressão linear) e pode ser representado por um sistema de equações. Um modelo neural é representado como uma arquitetura (e.g. rede *feedforward*) de nós e conexões (com pesos) entre eles, além de uma função de aprendizado. Já os modelos simbólicos são geralmente representados por regras do tipo IF-THEN ou árvores de decisão. Enquanto os modelos de dependências típicos, cujo maiores expoentes são redes bayesianas, estabelecem uma estrutura de relacionamento causal entre as variáveis aleatórias de um domínio de

aplicação. Por fim, os modelos baseados em algoritmos genéticos têm sido aplicados nas operações de classificação e otimização. Uma vasta gama de algoritmos referenciados na literatura que seguem esses modelos, por exemplo, em (Quinlan, 1993); (Heckerman, 1997); (Haykin, 2001) e (Lopes et al., 1999). Além disso, são amplamente utilizadas soluções híbridas, as quais combinam os modelos supracitados com vistas à otimização do processo de extração de conhecimento em termos, por exemplo, do aprendizado, da estimativa de parâmetros e da melhoria da interpretabilidade dos padrões descobertos (Goldschmidt e Passos, 2005).

A escolha do melhor algoritmo para MD é freqüentemente crítica, pois é sabido que nenhum deles tem desempenho ótimo em todos os domínios de aplicação (Salzberg, 1997). A seleção desses algoritmos é realizada pelo analista e deve ser pautada nas restrições do domínio e/ou nas preferências do usuário final (e/ou especialista no domínio). Considerando essas restrições, o analista pode selecionar o algoritmo baseado em alguns parâmetros como, por exemplo, o tipo de aprendizado, paradigmas de aprendizado, linguagens de representação e desempenho.

Vale ressaltar que, além da observação desses parâmetros, as avaliações experimentais desempenham um papel fundamental na seleção de um algoritmo, uma vez que não existem métodos formais para decidir qual o melhor algoritmo para um determinado domínio de aplicação (Salzberg, 1997).

2.2.4. Avaliação do Conhecimento Extraído (pós-processamento)

O processo de MD não termina quando os padrões nos dados de entrada são extraídos. É preciso realizar ainda uma etapa, comumente chamada pós-processamento, a qual deve permitir que o usuário entenda e possa julgar a utilidade do conhecimento extraído, contrastando-o com o conhecimento do especialista do domínio. Essa interação pode facilitar a busca das causas de possíveis erros ocorridos ao longo de todo esse processo.

O êxito obtido nesse processo está intimamente relacionado à facilidade com que o usuário pode compreender os padrões descobertos. Essa compreensibilidade é freqüentemente decorrente do modo como o conhecimento está organizado (representado) e de como realizar inferências sobre esses padrões.

Para aplicações reais, é fundamental que as inferências possam ser realizadas mesmo em situações que envolvem incerteza. Segundo Turban (2001), as técnicas de inteligência computacional tratam incerteza como um processo contendo três passos básicos, conforme apresentado na figura 2.4.

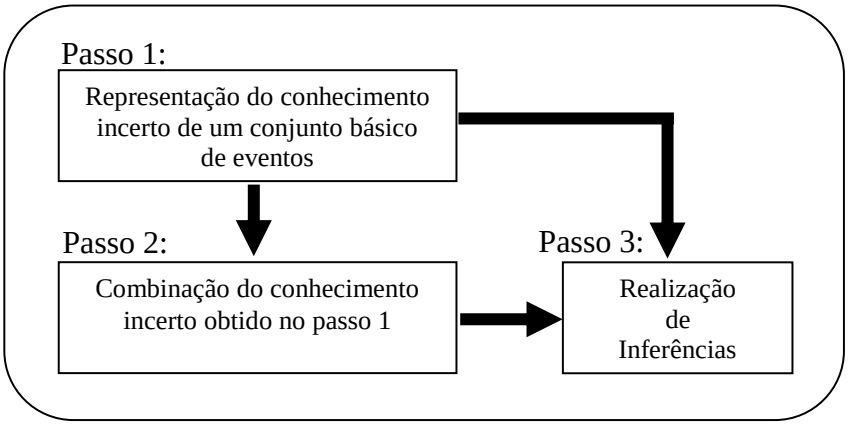

Figura 2.4. Os três principais passos para a manipulação de incerteza com técnicas de inteligência computacional.

No passo 1, o conhecimento inexato é obtido, por meio de interações com o especialista de um domínio específico e/ou induzido a partir de dados, e representado geralmente de modo numérico (e.g. valores de probabilidade), gráfico (e.g. diagrama de influência) ou simbólico (e.g. regras fuzzy). O conhecimento básico modelado no passo 1 pode ser diretamente usado para a realização de inferências (passo 3). Entretanto, na maioria dos casos reais, é necessário combinar os padrões obtidos no passo 1 a partir do uso de robustas técnicas de raciocínio incerto, tais como sistemas fuzzy e redes bayesianas. Por fim, o passo 3 funciona como uma máquina de inferência para induzir o conhecimento obtido a partir dos passos 1 e 2.

Desse modo, para que o fluxo de execução do processo mostrado na figura 2.4 possa ser realizado a contento, é necessário prover modelos cada vez mais confiáveis e de fácil interpretação do conhecimento extraído, de tal sorte que seja possível, para o usuário, identificar o quanto esse conhecimento é interessante e relevante para o suporte de seus processos decisórios.

A avaliação do modelo (padrões) é uma tarefa bastante difícil que envolve, entre outros aspectos, a utilização de medidas objetivas e subjetivas. Quanto à primeira, devem-se considerar critérios como a precisão, tempo de aprendizado, cobertura e suporte. Já

quanto às medidas subjetivas, é importante sublinhar aspectos como o grau de interesse (interessabilidade) do modelo (Rezende, 2003).

Medidas subjetivas são relevantes para aferir não a estrutura do modelo descoberto (e.g. a cobertura e confiabilidade das regras de decisão), avaliada pelas medidas objetivas, mas para quantificar o grau de interesse do conhecimento obtido para o usuário, a partir do modo como cada usuário considera os padrões descobertos. Isto é, reconhecer que o interesse de um padrão para um usuário não necessariamente é o mesmo para outro usuário.

Esse grau é medido com base nas expectativas do usuário sobre os padrões a serem descobertos e o conhecimento prévio dos dados. Desse modo, o grau de interesse de um padrão descoberto pode ser definido em termos da crença que o usuário tem sobre e existência do mesmo e, em consequência, o grau de interesse desse padrão é medido pela influência que o mesmo exerce sobre a crença do usuário. Para medir essa influência é considerado, frequentemente, se um padrão é "útil" e/ou "inesperado".

Um padrão é útil se o usuário pode se beneficiar, tirar algum proveito dele. Essa medida é investigada de modo detalhado em (Piatetsky-Shapiro e Matheus, 1994). Já a "inesperabilidade" de um padrão decorre da surpresa que esse padrão representa para o usuário. Essa abordagem, utilizada nesta tese, é discutida com propriedade em (Silberschatz e Tuzhilin, 1996).

No contexto da medida do grau de interesse de um padrão, com base em sua capacidade de ser inesperado, é possível estabelecer dois tipos de crenças que esse usuário pode ter sobre modelo de conhecimento descoberto (Silberschatz e Tuzhilin, 1996):

- *Crença forte*: a crença não é modificada com novas evidências. Caso uma nova evidência contradiga essa crença, então o usuário acredita que é muito provável que haja um erro na obtenção dessa crença (e.g. a partir de dados incorretos). Por exemplo, caso haja uma evidência de que o número de eleitores de uma cidade é maior que o número de habitantes dessa cidade, então a mesma se torna bastante questionável.
- *Crença fraca*: nesse tipo de crença o usuário está disposto a modificar sua opinião (crença) a cerca de um padrão, de acordo com as novas evidências. Por

exemplo, o usuário pode acreditar que o número de eleitores, da classe econômica A, do candidato C_1, é maior que o número de eleitores, da mesma classe, para o candidato C_2, até que uma nova evidência (e.g. nova pesquisa) seja apresentada para esse usuário. Para cada crença fraca pode ser assinalado um grau que especifica o quão forte é essa crença, na avaliação do usuário.

Para realizar as medidas de crença sobre os padrões, podem ser utilizadas diversas abordagens, como a bayesiana, *Dempster-Shafer* e frequência. O grau de interesse de um padrão é calculado em função da influencia do mesmo sobre a crença do usuário, medida por meio de uma dessas abordagens. Isto é, quanto maior for o impacto desse padrão sobre a crença do usuário, maior deve ser o grau de interesse desse padrão (Silberschatz e Tuzhilin, 1996).

Por fim, cabe salientar que os padrões inesperados são, via de regra, também considerados padrões úteis por parte dos usuários. Essa forte relação entre essas medidas fazem como que as abordagens que utilizam medidas de inesperabilidade sejam também utilizadas como medida de utilidade de um padrão e vice-versa (Malhas e Aghbari, 2009) e (Silberschatz e Tuzhilin, 1996).

2.2.5. Consolidação e Utilização do Conhecimento Extraído

A consolidação do conhecimento extraído pressupõe a verificação e a solução de potenciais conflitos com o conhecimento previamente extraído antes de o processo ser iniciado.

O conhecimento pode, então, ser organizado pelo analista dentro de um novo modelo, usado para refinar o modelo existente na aplicação ou simplesmente documentado e informado ao usuário.

2.3. PRINCIPAIS DESAFIOS E TENDÊNCIAS NA ÁREA DE MINERAÇÃO DE DADOS

Por ser uma tecnologia relativamente recente, há ainda muitos desafios e interessantes assuntos a serem investigados em MD. São apresentadas a seguir algumas dessas

perspectivas e pontos que vem merecendo cada vez mais a atenção da comunidade científica (Rezende, 2003); (Tan et al., 2005); (Goldschmidt e Passos, 2005); (Han e Kamber, 2006); (Stankovskia et al., 2008):

a) Estabelecimento de critérios de avaliação das diferentes abordagens dos algoritmos de MD, tais como:

   - Escalabilidade: número de registros e atributos, múltiplas relações *versus* tempo de aprendizagem;
   - Robustez: tolerância a ruídos e dados incompletos, atributos categóricos com um elevado número de estados;
   - Compreensibilidade: medidas de interesse para avaliar os padrões descobertos;
   - Parametrização: estabelecimento de métrica para aferir o impacto de determinados ajustes nos parâmetros dos algoritmos no resultado do processo de mineração.

b) Mineração interativa de conhecimento em níveis múltiplos de abstração e incorporação de conhecimento de fundo: a informação referente ao domínio estudado deve ser usada para guiar o processo de descoberta e permitir que os padrões descobertos sejam expressos em termos concisos e em níveis diferentes de abstração.

c) Linguagens de consulta de mineração de dados e mineração de dados *ad hoc*, a partir do desenvolvimento de linguagens de consultas para MD, preferencialmente integradas em linguagens de consulta de banco de dados ou de *data warehouse* e otimizadas para a mineração de dados eficiente e flexível.

d) Apresentação e visualização dos resultados da MD: o sistema deve apresentar requisitos de expressão do conhecimento descoberto em linguagens de alto nível, representações visuais ou outras formas expressivas, em termos humanos.

e) Suporte a ruídos ou a dados incompletos: o sistema deve prover métodos de limpeza e análise para o tratamento de ruídos e dados incompletos, os quais

comprometem a precisão dos padrões descobertos. Devem também prover métodos de mineração de *outliers*, para a descoberta e análise de exceções.

f) Suporte de tipos de dados relacionais e complexos: suporte a dados complexos, como hipertexto, multimídia, dados espaciais, temporais ou transacionais.

g) Sistemas de *DataMining Grid*, considerando caraterísticas críticas como flexibilidade, extensibilidade, escalabilidade, eficiência e facilidade de uso.

h) Mineração em bancos de dados heterogêneos e distribuídos em sistemas de informações globais: valendo-se da intra e internet, esses sistemas possuem suas bases de dados constituídas de muitas fontes, altamente distribuídas e heterogêneas (dados semiestruturadas e não-estruturadas, tais como em documentos de texto e da Web). A MD pode ser empregada para descobrir padrões em bases de dados heterogêneas e melhorar a interoperabilidade nesse tipo de base. Desse modo, pesquisas relacionadas à descoberta de conhecimento, usando Web Mining, ganha destaque, principalmente no que concerne à descoberta de padrões de uso, conteúdo e de estrutura na Web.

É possível destacar, ainda, alguns outros temas relevantes como proteção da privacidade e segurança da informação em MD, mineração de dados ubíqua, mineração visual de dados e interação humano-computador, incluindo o emprego de ferramentas de visualização de dados e conhecimento, além da elaboração de interfaces de extração de padrões amigáveis, de tal modo que possibilite a maior interatividade possível ao processo de MD.

2.4. CONSIDERAÇÕES FINAIS

Neste capítulo, foram descritos os fundamentos básicos do processo de MD que dão suporte aos processos de decisão, com destaque para os requisitos e as ações comumente realizadas nas etapas que compõe esse processo.

Para realização da extração de padrões em MD, é muito comum o uso de técnicas para raciocínio incerto com o objetivo de criar modelos do conhecimento embutido nos dados, em domínios reais com alguma inexatidão em suas premissas, e, assim, dar suporte

ao processo de decisão relacionado às ações futuras sobre esse domínio. Apesar das técnicas de manipulação de incerteza não estarem exclusivamente associadas à resolução de problemas de mineração de dados, são mecanismos eficientes e amplamente utilizados para análise de dados. No próximo capítulo, os principais conceitos dessas técnicas são investigados, além de como os mesmos podem ser empregados em MD.

3. TÉCNICAS DE RACIOCÍNIO INCERTO PARA MINERAÇÃO DE DADOS

3.1. CONSIDERAÇÕES INICIAIS

As tarefas de MD podem ser realizadas via técnicas para a manipulação de incerteza. Não obstante essas técnicas poderem ser utilizadas para diversos outros propósitos, inclusive para a resolução das tarefas de MD, como classificação, regressão e modelagem de dependência. Na figura 3.1, é apresentada uma integração das principais tarefas de mineração com os mecanismos básicos de manipulação de incerteza, compondo o objetivo primordial de ambas – a extração de conhecimento (Chen, 2001).

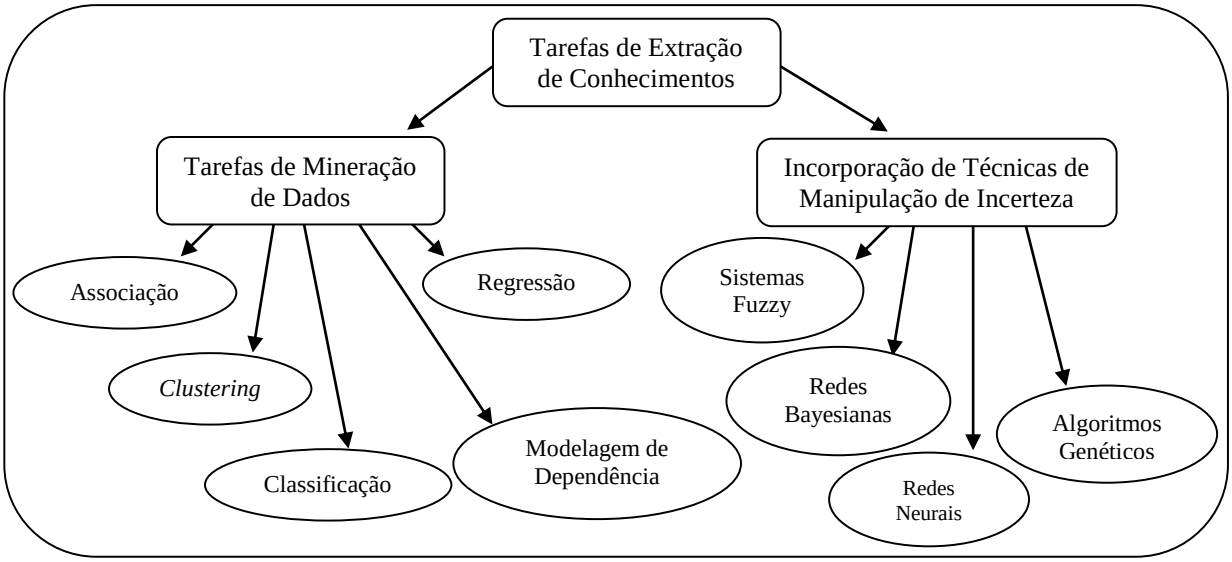

Figura 3.1. Tarefas de Extração de Conhecimento. Adaptada de (Chen, 2001).

Por questões de afinidade com os temas tratados nesta tese, são apresentadas a seguir as RBs e os algoritmos genéticos.

3.2. ALGORITMOS GENÉTICOS

Um Algoritmo Genético (AG) é um método de otimização e busca inspirado em mecanismos de populações de seres vivos, foi introduzido por John Holland, em 1975, e popularizado por um de seus alunos, David Goldberg, em 1989 (Goldberb, 1989). Esses algoritmos destacam-se por possuírem amplos espaços de busca e por serem do nível de complexidade de problemas NP-Completos.

Os AGs seguem o princípio da seleção natural e sobrevivência do mais apto e se baseiam no trabalho de Charles Darwin. Os AGs buscam a melhor solução para um dado problema, tentando várias soluções e utilizando-se das informações do próprio domínio para encontrar soluções cada vez melhores.

Os AGs têm sido empregados em problemas complexos de otimização fundamentalmente em virtude dos seguintes fatores:

- Funcionam tanto com parâmetros contínuos como discretos, ou ainda com uma combinação dos mesmos;
- Realizam buscas simultâneas em várias regiões do espaço de busca, pois trabalham com uma população e não com um único ponto;
- Utilizam informações de custo ou recompensa e não derivadas ou outro conhecimento auxiliar;
- Otimizam parâmetros de funções objetivos com superfícies complexas, reduzindo a incidência de mínimos locais;
- Trabalham com uma codificação do conjunto de parâmetros e não com os próprios parâmetros;
- Fornecem uma lista de parâmetros ótimos e não uma simples solução;
- São fáceis de serem implementados em soluções computacionais;
- São modulares e portáteis, no sentido que o mecanismo de evolução é separado da representação particular do problema considerado. Assim, eles podem ser transferidos de um problema para outro;

- São também facilmente combinados com outras técnicas heurísticas.

As técnicas de AG resolvem problemas de otimização e de busca, onde os métodos tradicionais falham. As técnicas tradicionais, frequentemente, iniciam o processamento com um único candidato (indivíduo) manipulado, utilizando alguma heurística, na maioria das vezes estática, diretamente ligadas ao problema a ser solucionado. No caso de AG, são realizadas operações em paralelo sobre uma população de candidatos (vários indivíduos) e a busca é feita, em diferentes áreas do espaço de solução, selecionando um número apropriado de membros para a busca, em várias regiões.

A idéia básica do funcionamento de um AG pode ser resumida no passos apresentados na figura 3.2:

1. Gerar a população inicial de cromossomos.
2. Avaliar cada cromossomo e lhe atribuir uma nota (aptidão).
3. Selecionar os melhores indivíduos (cromossomos mais aptos).
4. Aplicar operações de *crossover* (combinação) e mutação, gerando uma nova população (descendentes).
5. Repetir dos itens 2 a 4 até encontrar uma solução satisfatória.

Figura 3.2. Passos básicos de um Algoritmo Genético.

Conforme apresentado na figura 3.2, a base do funcionamento de um AG são os operadores genéticos, utilizados para gerar novas soluções (gerações da população) em analogia à evolução natural. Essas operações bem como a terminologia empregada nesse método de otimização, serão descritas nas subseções a seguir.

3.2.1. Terminologia

Um Algoritmo Genético representa uma metáfora da teoria da evolução, permitindo o uso de termos próprios da biologia. Desta forma, para uma melhor compreensão das relações destes termos com a teoria dos AGs, os mesmos são descritos a seguir:

- Cromossomo e Genoma: genoma é o conjunto completo de genes de um organismo. Um genoma pode ter vários cromossomos. Ambos representam a estrutura de dados para codificar uma solução do problema proposto. Um cromossomo ou genoma representa um ponto no espaço de busca;

- Gene: unidade de hereditariedade transmitida pelo cromossomo, o qual controla as características do organismo. Nos AGs, é um parâmetro codificado no cromossomo, um elemento do vetor cromossomo;
- Indivíduo: um simples membro da população. É formado pelo cromossomo e sua aptidão;
- Genótipo: é a composição contida no genoma. Na Computação Evolucionária, representa a informação contida no cromossomo ou genoma;
- Fenótipo: representa o objeto, estrutura ou organismo construído pelas informações contidas no genótipo (cromossomo decodificado);
- Alelo: é uma das formas alternativas de um gene. Em AG, representa os valores do parâmetro codificado em um gene.

3.2.2. Representação

A representação de um cromossomo pode conter várias formas, sendo a mais usual a representação binária, na qual as informações são codificadas em grandes cadeias de bits, como forma de utilizar os operadores de *crossover* e mutação tradicionais. Entretanto, é possível empregar outras formas de representação, dependendo do tipo de aplicação desenvolvida.

Vários pesquisadores têm discutido qual a melhor representação e muitos deles têm mostrado experimentos favoráveis à representação real (Michalewicz, 1994), principalmente por ser mais facilmente compreendida pelo ser humano. Todavia, a representação binária é mais simples de ser implementada, além de ser amplamente utilizada para a representação dos cromossomos (possíveis soluções de um problema). A escolha do tipo de representação é pautada, via de regra, pelo tipo de problema que está sendo modelado.

3.2.3. População Inicial

Os parâmetros de entrada de um AG incluem uma população inicial de cromossomos. Estes representam as estruturas de dados usados para representar uma possível solução do problema proposto. Cada cromossomo se encontra no espaço de busca de soluções, o qual representa o conjunto de todas as configurações que um cromossomo pode assumir. A cada cromossomo é atribuída uma aptidão, ou seja, uma nota ou valor para medir a qualidade da solução codificada.

Quando não existe um conhecimento a priori do espaço de busca, usa-se uma população inicial aleatória, porém, uma população inicial pequena pode não representar todas as regiões do espaço de busca. As possíveis soluções para este problema incluem (Goldberg, 1989):

- Gerar uma população inicial uniforme (com pontos espaçados igualmente entre si), distribuída por todo o espaço de busca;
- Gerar metade da população aleatoriamente e a outra metade invertendo os bits da primeira;
- Gerar uma população inicial maior que as subseqüentes, melhorando a representação do espaço de busca;
- Utilizar a técnica denominada *seeding*, que consiste em colocar na população inicial, soluções encontradas por outros métodos de otimização, garantindo soluções melhores para o AG que as geradas por esses outros métodos.

3.2.4. Avaliação e Seleção

Após a criação da população inicial, o AG seleciona os cromossomos com melhor aptidão. Os cromossomos mais aptos são os escolhidos para gerar os cromossomos filhos e são denominados de população intermediária. Indivíduos mais aptos têm mais oportunidades de serem reproduzidos (produzindo descendentes cada vez mais aptos).

A determinação da aptidão varia de acordo com o problema em questão. Por exemplo, caso se deseje determinar o máximo global de uma função complexa com vários

pontos de máximo e mínimo, pode-se utilizar o valor da função objetivo, interpolando estes valores para um intervalo determinado.

Os cromossomos podem ser escolhidos com probabilidade proporcional a sua aptidão, implicando na possibilidade de existência de indivíduos (cromossomos) duplicados na população intermediária. Esse procedimento, conhecido como “Método da Roleta” (ou "Roda da Roleta"), funciona do seguinte modo: com os cromossomos ordenados por aptidão, calculam-se as aptidões acumuladas, gerando um número aleatório *r* entre 0 (zero) e a soma total das aptidões. O cromossomo escolhido será o primeiro que possuir aptidão acumulada maior que *r*, como pode ser visto no exemplo ilustrado na tabela 3.1. Repete-se o processo até ser atingido o número de cromossomos desejados na população intermediária.

Tabela 3.1. Ordenação da aptidão interpolada para o intervalo [0,00; 2,00].

Posição	**Aptidão**	**Aptidão Acumulada**
1	2,00	2,00
2	1,50	3,50
4	1,00	4,50
5	0,50	5,00
6	0,00	5,00

A utilização deste método de seleção não trabalha com números negativos, sendo necessário o escalonamento dos valores das aptidões em um intervalo positivo, para tal utiliza-se a equação 3.1.

$$f_i = Min + (Max - Min) * \frac{(N - i)}{(n - 1)} \qquad \text{Equação 3.1}$$

Ao serem plotados em um gráfico, os dados da equação 3.1 se apresentam como mostrados na figura 3.3. Por este método, a probabilidade de escolha de cada cromossomo pode ser calculada pelo intervalo da aptidão acumulada. Enquanto o cromossomo 1 será escolhido para qualquer número *r* entre [0; 2[, o cromossomo 2 só será selecionado para valores de *r* entre [2; 3,50[, ou seja, dividindo-se o intervalo de 2 do cromossomo 1 pelo total de 5 obtemos 0,4, enquanto o cromossomo 2, obtém probabilidade de 1,5/5 = 0,3, e assim sucessivamente até o último cromossomo.

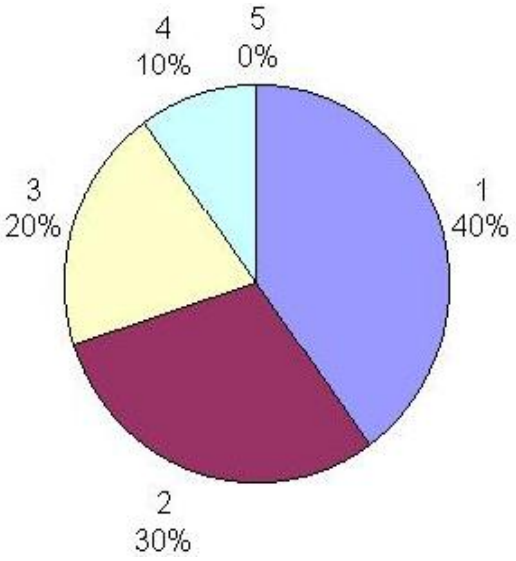

Figura 3.3. Distribuição das probabilidades da aptidão dos indivíduos.

Na figura 3.3, pode-se observar o fator de probabilidade agindo sobre a escolha dos indivíduos da população, enquanto o cromossomo 1 tem 40% de chance de ser escolhido, por possuir um intervalo maior, o cromossomo de posição 5 nunca será escolhido, pois se $4{,}50 < r <= 5{,}00$ será escolhido o cromossomo 4 e não o cromossomo 5.

Deve-se notar que o procedimento Roda da Roleta tende a causar convergência prematura do AG, pois a população intermediária pode ficar saturada de elementos com alta aptidão, mas não com a melhor aptidão, resultando em soluções não tão boas quanto o esperado.

O problema de convergência prematura pode ser combatido utilizando, principalmente, os seguintes expedientes:

- Aumentar da taxa de mutação. A diversidade é proporcional à taxa de mutação, porém, a mutação é conhecida por destruir informação genética dos cromossomos, por isso é necessário cautela quando do ajuste da taxa de mutação;
- Evitar cromossomos duplicados na população;
- Controlar o número de filhos do superindivíduo (indivíduo com alta aptidão, mas não com aptidão ótima) usando, por exemplo, técnicas como seleção por torneio.

A Seleção por Torneio evita o problema de convergência prematura. No torneio, escolhe-se aleatoriamente e com probabilidades iguais, *n* cromossomos e o de maior

aptidão entre eles é selecionado para a população intermediária. O processo é repetido até se preencher a população intermediária. Na seleção por torneio, não existe a necessidade de escalonamento da aptidão ou seu ordenamento, o que acarreta uma economia de recursos e tempo de processamento.

3.2.5. Operador de *Crossover*

O operador genético *crossover* é responsável pela geração de novas populações. Esse operador é aplicado a um par de cromossomos da população intermediária para gerar descendentes, o que permite a exploração do espaço de busca das soluções do problema tratado.

A idéia do *crossover* é mesclar em um novo cromossomo as melhores qualidades de dois indivíduos. Levando em consideração que as melhores características dos indivíduos não são conhecidas, o melhor a fazer é combinar as características aleatoriamente. O AG trata estas características como blocos de construção e tenta combiná-los em melhores indivíduos via *crossover*. Algumas vezes, o *crossover* combina as piores qualidades dos indivíduos e, neste caso, seus descendentes não sobreviverão muito tempo na população. O melhor caso, no entanto, consiste em combinar as melhores qualidades dos pais, gerando um filho com aptidão ainda melhor que a de seus progenitores, desde que essas características sejam compatíveis entre si.

O operador de *crossover* divide as cadeias de bits dos cromossomos pais em uma posição aleatória, produzindo um par de "cabeças" e outro de "caudas", a permuta de "caudas" dessas cadeias de bits entre os cromossomos pais dá origem aos cromossomos filhos, como pode ser visto na figura 3.4.

pai_1	***1011000***1101011	***1110001***0110110
pai_2	0010011***0100101***	0011100***0110011***
$filho_1$	***101100011100101***	***111000110110011***
$filho_2$	001001101101011	001110010110110

Figura 3.4: Exemplo da aplicação do operador de *crossover*.

O operador de *crossover* não ocorre em todos os pares selecionados. A probabilidade de aplicação do operador de *crossover* pode variar de acordo com uma taxa, conhecida como taxa de *crossover*, geralmente entre 60% e 99%. A não ocorrência de *crossover* implica na cópia dos cromossomos pais para a próxima geração, preservando algumas soluções da geração anterior. Uma implementação possível para a taxa de *crossover* poderia usar números pseudo-aleatórios no intervalo [0, 1], aplicando o *crossover* caso o número gerado seja menor que a taxa de *crossover*.

Os tipos do operador de *crossover* são determinados pelo número de pontos de corte utilizados, sendo os mais conhecidos: *crossover* de um-ponto, *crossover* multiponto e *crossover* uniforme. O *crossover* de um ponto divide as cadeias de bits dos cromossomos pais em um ponto escolhido aleatoriamente. O *crossover* de *n* pontos escolhe *n* pontos aleatórios nas cadeias de bits dos pais, onde *n* é um número par, e as duas seções entre estes dois pontos são permutadas. O *crossover* multiponto mais utilizado tem sido o de dois pontos, pois estes tendem a preservar esquemas[1] de maior comprimento. Um exemplo do *crossover* de dois pontos pode ser visto na figura 3.5.

pai_1	101**100**011	010**111**110	011**011**011
pai_2	001001101	0010100111	001001001
$filho_1$	101001011	0010100111	011001011
$filho_2$	001**100**101	0101111100	001**011**001

Figura 3.5. Exemplo de aplicação do *crossover* de dois pontos.

O *crossover* uniforme utiliza uma máscara de bits aleatórios gerada para cada par de cromossomos escolhidos para gerar novos filhos. Este processo pode ser implementado da seguinte forma: se o bit da máscara possui valor 1, o bit correspondente do pai_1 é copiado para o $filho_1$, caso contrário, o bit do pai_2 é copiado para o bit do $filho_1$. Este processo se repete por toda a cadeia de bits da máscara. Para a geração do $filho_2$, este processo é invertido. Na figura 3.6, é ilustrado o uso do *crossover* uniforme.

[1]Esquema é a identificação de certas cadeias de bits comuns a cromossomos de alta aptidão e podem ser utilizados para melhorar a aptidão média da população.

Máscara de bits	0110110110	1011000111	0010110011
pai_1	1011000111	0101111100	0110110110
pai_2	0010011011	0010100111	0010110011
$filho_1$	0010001111	0001100100	0010110011

Figura 3.6. Exemplo de aplicação do *Crossover* Uniforme.

Enquanto o *crossover* de *n* pontos gera um filho com metade dos bits de cada pai, o *crossover* uniforme gerará um filho, no qual seus genes terão um número aleatório de bits do pai_1 e do pai_2, devido à utilização da máscara aleatória de bits.

Vários estudos têm investigado a diferença de desempenho para os diversos tipos de *crossover* (Eshelman et al., 1989); (Beasley et al., 1993) e (Grefenstette, 1986). O que se pode deduzir é que não há grandes diferenças, pois a robustez do AG não apresenta alteração significativa em seu desempenho em uma faixa relativamente larga de variação de parâmetros (taxa de *crossover* e mutação, tamanho da população, etc).

3.2.6. Operador de Mutação

O operador de mutação é utilizado após o *crossover* e é considerado o melhor mecanismo para produzir variações nas características dos indivíduos da população. A mutação é aplicada, com certa probabilidade, a cada bit dos cromossomos filhos, invertendo o valor do bit sobre o qual é aplicado, criando novas tentativas de solução com pequenas mudanças aleatórias na representação anterior. Estas perturbações, nas cadeias dos cromossomos, darão origem a uma nova cadeia, evitando que a busca fique estagnada em sub-regiões do espaço de busca, possibilitando, também, que qualquer ponto do espaço seja atingido (Rezende, 2003).

Este operador contribui para aumentar a diversidade de cromossomos na população, entretanto, destrói informação contida no cromossomo. Dessa forma, deve-se utilizar uma taxa de mutação pequena, mas suficiente para assegurar a diversidade, normalmente entre 0,1% e 5%. Na figura 3.7, é apresentado um exemplo do uso deste operador.

Filho1	10110**0**01110101100**1**110010110011
Filho2	00100110110010111100011**1**10110
Filho1	10110**1**01110101100**0**110010110011
Filho2	00100110110010111100011**0**10110

Figura 3.7. Exemplo da aplicação do operador de mutação.

Após a aplicação dos operadores de *crossover* e mutação, os cromossomos filhos são então avaliados e o processo se repete em ciclos. Estes ciclos são chamados de gerações. O número de repetições (ou gerações) pode ser pré-determinado; indicado pela ocorrência do caso ideal (critério de parada), se este for conhecido; ou na convergência do algoritmo, ocorrendo quando as aptidões dos indivíduos de uma população se tornam muito parecidas.

A substituição dos cromossomos da população atual pode ocorrer de forma geracional ou *steady-state*. A primeira substitui toda a população, criando N filhos para substituir *N* pais, enquanto que a última cria um conjunto com os melhores pais e filhos para a próxima geração, neste tipo de substituição são criados um ou dois filhos em cada geração para substituir os piores cromossomos da população. Alternativamente podem-se substituir os pais ou os indivíduos mais velhos, pois estes já transmitiram seus genes à população.

3.2.7. Algoritmos Genéticos e Mineração de Dados

A utilização de AG em soluções baseadas em MD, em sua grande maioria, não foge à regra básica de aplicação dessa técnica, ou seja, na resolução de problemas de otimização. Entretanto, em (Lopes et al., 1999), é proposto o método *Rule Evolver,* baseado na teoria de AGs, que pode ser aplicado para solucionar problemas que envolvem classificação e sumarização.

Negoita (2000) identifica os AGs como um importante veículo de otimização para os processos de Mineração e levanta uma dicotomia. Por um lado, o esforço computacional dos métodos evolutivos é por vezes proibitivo. De outro, o aspecto estrutural do processo

de otimização suportado por esses métodos é bastante atrativo, particularmente em ambientes muito heterogêneos nos quais a variedade de padrões a serem explorados e eventualmente descobertos em bases de dados são altamente complexos.

Goebel e Gruenwald (1999) afirmam que os AGs são freqüentemente empregados para formularem hipóteses sobre as dependências entre as variáveis, na forma de regras de associação ou de outros formalismos internos.

Do mesmo modo que os sistemas fuzzy e as redes neurais, os AGs podem ser combinados com os demais métodos de manipulação de incerteza para as formulações inerentes à MD. Por exemplo, é possível integrar solução fuzzy com um AG, de tal forma que esse último possa ser usado para atualizar ou aperfeiçoar um sistema fuzzy. Um modo de realizar essa tarefa pode ser construído utilizando o AG para ajustar os parâmetros (e.g. regras) gerados inicialmente para esse sistema (Chen, 2001). Além disso, os AGs podem, aliados a outras técnicas, serem utilizados para a resolução de problemas próprios das etapas do processo de MD, principalmente na fase de pré-processamento (e.g. tratamento de ruído nos dados).

3.3. REDES BAYESIANAS

As RBs podem ser entendidas como modelos que codificam os relacionamentos probabilísticos entre as variáveis que representam um determinado domínio (Russel e Norvig 2003). Esses modelos possuem como componentes uma estrutura **qualitativa**, a qual representa as dependências entre os nós, e **quantitativa** (tabelas de probabilidades condicionais - TPCs desses nós), que avalia, em termos probabilísticos, essas dependências. Juntos, esses componentes propiciam uma representação eficiente da distribuição de probabilidade conjunta (DPC) do conjunto de variáveis aleatórias $X = \{X_1, X_2, \ldots, X_n\}$ de um determinado domínio (Pearl, 1988). Essa distribuição de probabilidade é dada pela equação 3.2:

$$P(x) = \prod_{i=1}^{n} P(x_i \mid pais(x_i)) \qquad \text{(Equação 3.2)}$$

na qual Pa_i são os nós-pais do nó X_i. Essa representação acarreta uma redução substancial do número de probabilidades a serem manipuladas, a partir da utilização do conceito de independência condicional, expressa por essa equação.

Uma RB ou rede de crença pode ser vista como um grafo que representa o relacionamento probabilístico entre o conjunto de variáveis de um domínio, consistindo em (Russell e Norvig, 2003):

- Um conjunto de nós que representam as variáveis aleatórias $X=\{X_1,..X_n\}$ do domínio;
- Um conjunto de ligações dirigidas ou setas conectando pares de nós. Essas setas representam a influência direta que um nó exerce sobre outro;
- Uma TPC para cada nó, que quantifica os efeitos que os nós-pais (*Pa*) exercem sobre esse nó. Os nós-pais de um nó *n* são todos aqueles que possuem setas apontando para *n*;
- Um grafo que não contém ciclos dirigidos. Daí dizer que são grafos acíclicos orientados ou dirigidos.

Desta forma, uma RB provê uma completa descrição do domínio. Cada entrada na DPC pode ser calculada a partir das informações da rede. Uma entrada genérica da DPC é a probabilidade de uma conjunção de uma particular atribuição de valores para cada variável, assim como $P(X_1 = x_1 \cap X_2 = x_2 ... \cap X_n = x_n)$. Utilizando a notação $P(x_1,...,x_n)$, os valores das entradas da DPC podem ser obtidos por intermédio da equação 3.2.

Cada entrada da distribuição de probabilidades conjunta é representada pelo produto dos elementos apropriados (de acordo com o relacionamento entre as variáveis) das tabelas de probabilidades condicionais na RB, ou seja, as TPCs estabelecem uma representação simplificada da DPC.

3.3.1. Independência Condicional

Para que a representação da DPC seja possível, a RB especifica, juntamente com as probabilidades condicionais da TPC, a independência condicional entre um subconjunto de

variáveis de X (Chen, 2001). Uma variável X (ou um conjunto de variáveis ($X_1,X_2,...,X_n$)) é condicionalmente independente de Y (ou de um conjunto de variáveis $Y_1,Y_2,...,Y_n$), dado uma variável Z (ou $Z_1,Z_2,...,Z_n$), se a distribuição probabilística de X é independente do valor de Y, dado o valor de Z, ou seja:

$$P(X|Y,Z) = P(X|Z)$$

Estendendo este conceito para um conjunto de variáveis, tem-se

$$P(X_1,X_2,...,X_n|Y_1,Y_2,...,Y_m,Z_1,Z_2,...,Z_k) = P(X_1,X_2,...,X_n|Z_1,Z_2,...,Z_k)$$

A definição de independência condicional corresponde à ausência de arcos entre os nós de uma RB. Por exemplo, os nós X e Y da rede visualizada na figura 3.8 são independentes condicionalmente, dado Z.

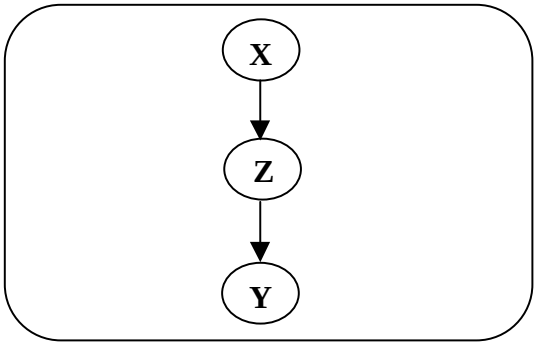

Figura 3.8. Exemplo de independência condicional.

A independência condicional entre as variáveis de um particular domínio pode auxiliar o processo de construção de uma RB para esse domínio. Uma forma de visualizar isso é baseada no seguinte:

- Reescrevendo $P(X_i)$ em termos da definição de probabilidade condicional e reduzindo cada conjunção de probabilidades a uma probabilidade condicional e uma conjunção menor, obtém-se:

$$P(x_1,x_2,...,x_n) = P(x_n,x_{n-1},...,x_1)P(x_{n-1}|x_{n-2},...,x_1) \ldots P(x_2|x_1)P(x_1)$$

Equação 3.3

- Comparando as equações 3.2 e 3.3, tem-se:

$P(x_i|x_{i-1},...,x_1) = P(x_i|Pa_i)$ Equação 3.4

Um outro modo de realizar o teste de independência é via medida de informação mútua, dada pela equação 3.5:

 Equação 3.5

Essa relação é reflexiva, $I(X,Y) = I(Y,X)$. Se $I(Y,X) > 0$, então X e Y são dependentes; em caso contrário, X e Y não são informativas entre si, ou seja, são independentes. Se X e Y são independentes, o argumento da função *log*, na equação 3.5, é igual a 1, para todos os valores de X e Y, o que leva o somatório ao valor 0 (zero).

3.3.2. Construção de Redes Bayesianas

A partir da equação 3.4, é possível notar que $Pa_i \subseteq \{x_{i-1},...,x_1\}$. Desse modo, para determinar a estrutura de uma RB deve-se (1) ordenar as variáveis de algum modo (geralmente de acordo com as observações do especialista do domínio) e (2) determinar o conjunto de variáveis que satisfazem a equação 3.4.

Para ilustrar o processo de construção de uma RB, considere o problema bastante simplificado da detecção de fraude em compras com cartões de crédito. Na figura 3.9, esse exemplo é mostrado, na qual os arcos são desenhados da causa para o efeito, os quadros mostram a distribuição de probabilidade local associada a cada nó da RB, e os asteriscos representam a atribuição de quaisquer valores, dentre os possíveis, às variáveis. Desse modo, a estrutura da RB pode ser obtida seguindo os passos:

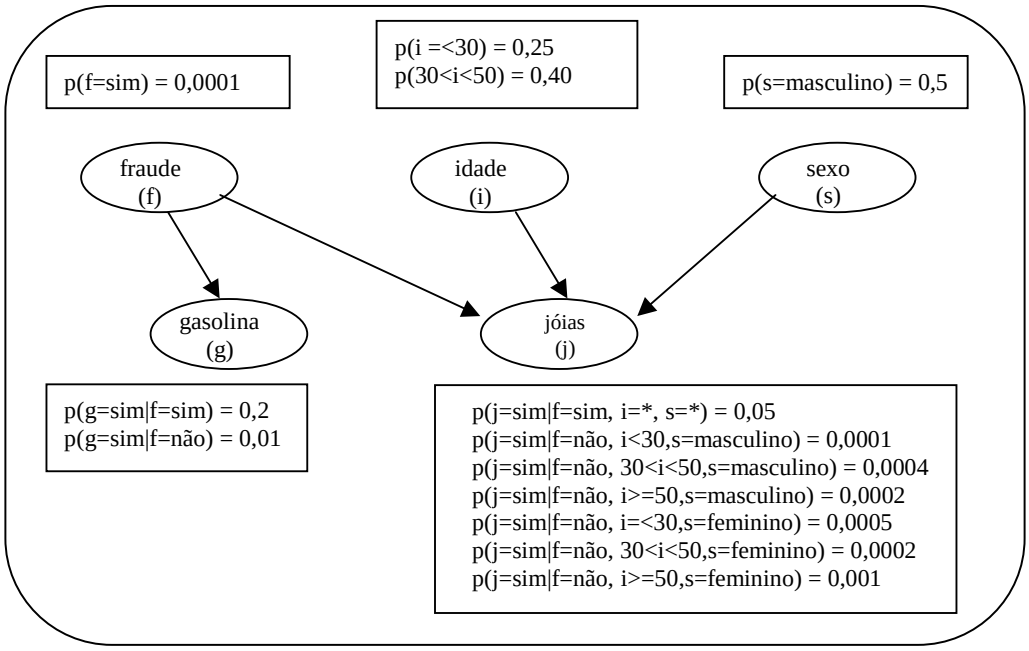

Figura 3.9. Exemplo de rede Bayesiana para detecção de fraude em compras com cartão de crédito (Heckerman, 1997).

1. Selecionar o conjunto de variáveis que descrevem o domínio. Uma possível escolha para o problema de detecção de fraude seria: *fraude(f)*, *gasolina(g)*, *jóias(j)*, *idade(i)* e *sexo(s)*, representando se a compra é ou não fraudulenta, se houve ou não compra de gasolina nas últimas 24 horas, se houve ou não compra de jóias nas últimas 24 horas, o sexo e a idade do usuário do cartão de crédito, respectivamente.

2. Eleger uma ordem para essas variáveis. Uma possível ordem das variáveis para o problema de detecção de fraude seria: *f*, *i*, *s*, *g* e *j*.

3. Enquanto houver variáveis a serem consideradas:

   a) Adicionar um nó à rede para cada variável.

   b) Estabelecer um conjunto de nós-pais de X_i sobre o conjunto mínimo de nós já incluídos na rede, que satisfazem a propriedade de independência condicional, estabelecida pela equação 3.4 ou 3.5. No exemplo, utilizando a ordem *f, i, s, g* e *j*, sugerida no passo 2, obtém-se:

$$P(i \mid f) = P(i)$$

$$P(s \mid f,i) = P(s)$$

$$P(g \mid f,i,s) = P(g \mid f)$$

$$P(j \mid f,i,s,g) = P(j \mid f,i,s)$$

c) Definir a TPC para $\boldsymbol{X_i}$.

Vale ressaltar que, apesar de serem mostrados como uma seqüência simples, na prática, esses passos são geralmente interligados e complexos. Por exemplo, julgamentos de independência condicional e/ou causa-efeito podem influenciar na construção da RB.

Além disso, esse processo de construção considera apenas o conhecimento prévio (ou de fundo) do domínio. Em outras palavras, a rede deve ser concebida por intermédio da interação com o especialista do domínio. Essa interação tem a finalidade de identificar os relacionamentos entre as variáveis de interesse, para em seguida codificá-los na rede. Por exemplo, o especialista no problema da detecção de fraude pode identificar a influência direta que a idade exerce sobre a aquisição de jóias e até mesmo quantificar essa dependência (através dos valores de probabilidades condicionais). Na próxima seção, é mostrado o aprendizado da estrutura e dos parâmetros (probabilidades) de uma RB a partir de dados.

3.3.3. Aprendizado de Redes Bayesianas

A aprendizagem de RBs consiste em induzir, a partir de um conjunto de dados, as distribuições de probabilidades condicionais e identificar as relações de independência existentes nesse conjunto. Esse processo de aprendizagem considera dois aspectos: aprendizagem da estrutura, quando não se tem *a priori* definido pelo especialista do domínio, tal estrutura; e a aprendizagem dos parâmetros, após a obtenção da estrutura, por intermédio das interações com o especialista ou induzida a partir dos dados.

Para um melhor entendimento desse processo, considere o exemplo de detecção de fraude em compras com cartão de crédito mostrado na figura 3.9. Primeiramente, a RB que representa as relações entre as variáveis desse problema é especificada, por exemplo, pelo usuário. Em seguida, é necessário que seja especificado como a distribuição de probabilidade de cada nó será representada. No caso do problema da detecção de fraude, as variáveis foram discretizadas em um número de estados (valores) para que cada distribuição de probabilidade possa ser representada em uma tabela (TPC), por exemplo, idade foi discretizada dentro dos valores (≤ 30, 30 - 50, ≥ 50).

Finalmente, o algoritmo tenta estimar as probabilidades (parâmetros) da TPC baseado no conjunto de dados de treinamento. Por exemplo, a célula *P(i≤30)* da TPC da variável *i* pode ser simplesmente calculada a partir do número de clientes do conjunto de dados de treinamento que tenham idade igual ou inferior a 30 anos. O parâmetro *P(j=sim|f=sim, i ≤ 30,s=feminino)* pode ser computado por intermédio da fração dos exemplos de treinamento onde a compra seja uma fraude, a idade do cliente seja igual ou inferior a 30 anos e o sexo seja feminino e que tenha comprado jóias nas últimas 24h. Tecnicamente estas são as estimativas de Máxima Verossimilhança de cada um dos parâmetros da RB.

Um ponto que merece ser destacado no processo de aprendizado de RBs é se todos os valores das variáveis são observados (não há valores de atributos ausentes) no conjunto de dados de treinamento ou se algumas não são consideradas. Assim, os métodos de aprendizagem devem considerar as seguintes situações:

- estrutura da rede conhecida e conjunto de dados completos;
- estrutura conhecida e conjunto de dados incompletos;
- estrutura desconhecida e conjunto de dados completos;
- estrutura desconhecida e conjunto de dados incompletos.

Em razão do estudo de caso empregado neste trabalho lançar mão sempre de conjuntos de dados completos, esses métodos são enfatizados.

3.3.4. Aprendizado das Probabilidades em Redes Bayesianas

No caso em que a estrutura da RB é dada, o aprendizado se resume à atualização das tabelas de probabilidades condicionais dos nós dessa rede, caso as variáveis que eles representem sejam discretas ou através de uma função de distribuição de probabilidade, no caso das variáveis serem contínuas (Buntine, 1996).

Para uma melhor compreensão do aprendizado das probabilidades das TPCs de uma RB qualquer, considere uma rede definida a partir de um conjunto de variáveis discretas $X=\{X_1,X_2,...,X_n\}$ e por um modelo (estrutura) *M* (fornecido *a priori*) de RB, que especifica

as dependências condicionais dos elementos de X. Considere ainda que cada variável X_i possua um conjunto finito de estados (valores) c_i e que podem possuir um conjunto de variáveis pais (nós-pais) Pa_i, as quais podem assumir valores q_i. O modelo M produz, a partir da equação 3.6, uma fatoração da DPC de um conjunto de valores (um exemplo do conjunto de dados) das variáveis de X ($x_k=\{X_{ik},...,X_{nk}\}$) do seguinte modo:

Equação 3.6

no qual pa_{ij} denota o estado de Pa_i em x_k.

Suponha agora que seja fornecido um conjunto de dados $D = \{x_i,...,x_n\}$ contendo n exemplos (casos). A tarefa básica de um método de aprendizado, dada uma estrutura de RB M, é estimar, a partir de D, as probabilidades condicionais das TPCs de cada um dos nós de M. Estas probabilidades condicionais serão consideradas como parâmetros desconhecidos $\theta=(\theta_{ijk})$ (no qual $\theta_{ijk}= P(x_{ik}|pa_{ij}, \theta)$) e $\theta ij=(\theta_{ij1},..., \theta_{ijci})$ é o vetor de parâmetros (probabilidades condicionais da TPC de um nó) associados com a distribuição condicional $X_i|pa_{ij}$, inferido a partir de D.

Na figura 3.10, é mostrado um exemplo de uma RB definida por $X=\{X_1,X_2,X_3\}$ e $c_i=2$ para i = 1,2,3. O grafo, dado *a priori*, codifica a independência condicional de X_1 e X_2 e determina que ambos sejam pais de X_3. Desta forma, Pa_3 possui quatro valores pa_{ij} correspondente as quatro combinações de estados de X_1 e X_2 (para simplificar, esse valores serão denotados como pa_{31} = (1,1), pa_{32} = (1,2), pa_{33} = (2,1), pa_{34} = (2,2)). Assim, seis parâmetros (valores das TPCs) independentes $\theta=(\theta_1, \theta_2, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34})$ quantificam a RB, no qual $\theta_{11} = P(x_{11}|\theta)$, $\theta_{21}=P(x_{21}|\theta)$, e $\theta 3j1=P(x31|pa3j, \theta)$ para j = 1, 2, 3, 4. A partir desses parâmetros e considerando que os valores para cada linha das TPCs são coletivamente exaustivos, cuja soma das probabilidades é igual a 1, os seguintes vetores de parâmetros são obtidos: $\theta_1 = (\theta_{11}, 1-\theta_{11})$, $\theta_2 = (\theta_{21}, 1-\theta_{21})$, e $\theta_{31} = (\theta_{311}, 1-\theta_{311})$, $\theta_{32} = (\theta_{321}, 1-\theta_{321})$, $\theta_{33} = (\theta_{331}, 1-\theta_{331})$, $\theta_{34} = (\theta_{341}, 1-\theta_{341})$.

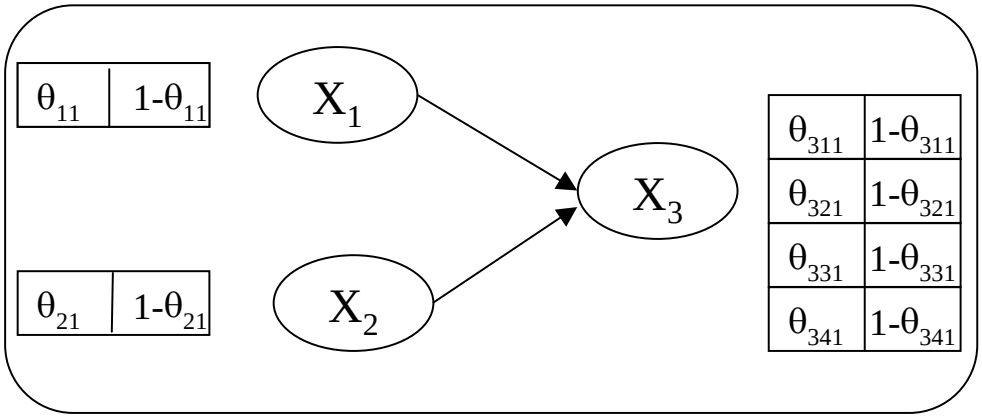

Figura 3.10. Parâmetros de uma rede Bayesiana.

Vários métodos de estimativas dos parâmetros θ_{ijk} podem ser encontrados na literatura. Dentre os mais utilizados, destacam-se o de Estimativas de Máxima Verossimilhança (*Maximum Likelihood*) e de Estimativas Bayesianas (Heckerman, 1997) (Gaag, 1996). As estimativas de máxima verossimilhança não consideram conhecimento a priori sobre as distribuições de probabilidade, utilizando-se apenas dos dados disponíveis. Estimativas Bayesianas utilizam os dados e algum conhecimento a priori expresso na forma de distribuições de *Dirichlet* para estimar os parâmetros a posteriori (Luna, 2004).

Como forma de compreender os processos de estimativa de parâmetros de uma RB, serão apresentados a seguir os principais básicos dos métodos de Estimativas de Máxima Verossimilhança e Maximum a Posteriori (Estimativa Bayesiana).

Considere $n(x_{ik}|pa_{ij})$ a freqüência de ocorrência de (x_{ik},pa_{ij}) em um conjunto de dados D, e $n(pa_{ij}) = \sum_k n(x_{ik}|pa_{ij})$ a freqüência de pa_{ij}. A distribuição de probabilidade condicional de um exemplo (caso) x_k de D pode ser expressa como uma função de θijk da seguinte forma:

$$P(x_k|\theta) = \prod_{i=1}^{n} \theta_{ijk}$$ Equação 3.7

Dessa forma, considerando os exemplos de D independentes, a distribuição de probabilidade conjunta de D é dada por

$$P(D|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}$$ Equação 3.8

A equação 3.8 é chamada de função de verossimilhança $l(\theta)$. A Máxima Verossimilhança representa, portanto, o cálculo dos valores dos parâmetros que maximizam $l(\theta)$. Geralmente, a estimativa de Máxima Verossimilhança para θijk é representado pela freqüência dos casos relevantes do conjunto de dados: $\hat{\theta}_{ijk} = \frac{n(x_{ik}|pa_{ij})}{n(pa_{ij})}$.

O método da Máxima Verossimilhança pode apresentar dois problemas muito comuns, em se tratando de algoritmos de aprendizado - dados esparsos e *overfitting*. Para entender estes problemas, considere um conjunto de dados *D* (referente à RB da figura 3.10), formado pelos exemplos mostrados na tabela 3.2.

Tabela 3.2. Conjunto de dados D.

Exemplo	X_1	X_2	X_3
1	V	F	V
2	V	V	V
3	V	V	F
4	V	F	F

O primeiro problema ocorre quando $\hat{\theta}_{ijk} = 0$ se $n(x_{ik}|pa_{ij}) = 0$, por exemplo, $P(X_3=V|X_1=F,X_2=V,\theta)$ não pode ser estimado, visto que não existem instâncias em *D* cujo $X_1 = F$. Segundo Buntine, para *n* variáveis binárias e uma RB totalmente conectada (como no exemplo da figura 3.10) são necessários mais do que 2^{n-1} exemplos em *D* para que a Máxima Verossimilhança possa ser estimada (Buntine, 1996). Uma possível solução para o problema da manipulação de poucos exemplos no conjunto de dados é ajustar a probabilidade de uma determinada célula de acordo com suas "adjacentes" na TPC. Por exemplo, pode-se definir que $P(X_3=V|X_1=F,X_2=V)$ tenha um valor próximo de $P(X_3=V|X_1=V,X_2=V)$ da TPC do nó X_3. Além desse método, vários outros são propostos na literatura (Dietterich, 1997).

Para entender o problema de *overfitting*, considere o cálculo da Máxima Verossimilhança para $P(X_1=V|\theta)$. O resultado será igual a *1,0*, pois em todos os exemplos $X_1 = V$. Uma solução para esse problema seria imaginar que, como esse parâmetro $P(X_1=V|\theta) = 1,0$ foi baseado em apenas 4 casos, parece razoável que o "verdadeiro" valor deva ser, por exemplo, em torno de *0,9*, e por acaso ocorreu de X_1 ser a igual a V em todos os exemplos. Portanto, a estimativa *1,0* deveria ser entendida como o limite superior de

probabilidade. Por definição, o valor de Máxima Verossimilhança ($1{,}0^4$) representa um valor superestimado da verdadeira verossimilhança dos dados ($0{,}9^4$). Quando isso ocorre diz-se que houve *overfitting* nos dados. De acordo com a Teoria da Máxima Verossimilhança, à medida que a amostra de dados vai aumentando a superestimação vai gradualmente convergindo para o "verdadeiro" valor da verossimilhança dos dados (Casella e Berger, 1990).

Um outro modo de calcular os parâmetros das TPCs de uma RB é conhecida como *Maximum a Posteriori* (MAP) e representa uma extensão da Máxima Verossimilhança. A MAP, que é um dos métodos mais utilizados pelos algoritmos de aprendizado de RBs, permite incorporar informações a priori (conhecimento de fundo) dos parâmetros θ_{ijk} antes dos exemplos de *D* serem vistos. Dessa forma, o aspecto central desse método é considerar θ_{ijk} como uma variável aleatória, cuja probabilidade a priori representa o grau de crença do observador (especialista) sobre as probabilidades condicionais das TPCs da RB, desconsiderando os dados. O método de aprendizado de redes Bayesianas então usa esses dados para atualizar as crenças a priori usando a regra de Bayes, da seguinte forma (Ramoni, 2001):

Equação 3.9

No qual *P(D)* é calculado conforme a equação 3.10:

Equação 3.10

Nesta seção, a estrutura da RB foi considerada a priori. Dessa forma, o aprendizado se resumiu ao cálculo das probabilidades das TPCs a partir dos dados. Na próxima seção, será considerado o processo de aprendizado da estrutura de uma RB.

3.3.5. Aprendizado da Estrutura de Redes Bayesianas

O aprendizado da estrutura de RBs a partir de dados completos pode ser realizado, via de regra, considerando dois paradigmas: busca e pontuação e baseado em independência condicional (Cheng et al., 1997).

No paradigma de busca e pontuação, a aprendizagem se realiza buscando uma estrutura que seja aderente aos dados. Em geral se inicia com um grafo sem arcos, então, usa-se algum método de busca gulosa que adicione um arco ao grafo. O passo seguinte consiste em usar uma função de pontuação para determinar se a nova estrutura é melhor que a anterior. Caso seja melhor, o novo arco é mantido. Esse processo continua até que nenhuma nova estrutura seja melhor que as anteriores.

Diferentes critérios de pontuação estão disponíveis na literatura para avaliar uma estrutura, tais como os descritos em (Cooper e Herskovits, 1992), (Lam e Bacchus, 1994) (Heckerman et al., 1995). Já o processo de busca por novas estruturas é realizado via métodos heurísticos. Para reduzir o espaço de busca, a maioria dos algoritmos requer ordenamento a priori dos nós.

Os algoritmos que utilizam o paradigma de análise de independência condicional procuram descobrir as dependências a partir dos dados, e então usam essas dependências para inferir a estrutura. As relações de dependência são avaliadas pelo uso de alguma classe de teste de independência condicional. Detalhes mais aprofundados destes tipos de algoritmos podem ser encontrados nos algoritmos descritos em (Spirtes, 2001) e (Cheng et al., 1997).

Como forma de compreender os processos de aprendizado da estrutura de uma RB e em razão de ser mais comumente utilizado, serão destacados a seguir os principais básicos do paradigma de busca e pontuação.

Para isso, é necessário entender que o aprendizado da estrutura de RBs pode ser visto como processo de busca por uma estrutura que codifique a DPC para um conjunto de variáveis aleatórias $\boldsymbol{X}$, dado um conjunto de dados D. Isto é, encontrar uma estrutura, em um possível espaço de hipóteses de estruturas S^h, avaliando as probabilidades a priori das hipóteses S^h - $P(S^h)$. Então, dado um conjunto de dados D, calcular as probabilidades condicionais que maximizam $P(S^h|D)$ (estrutura) e $P(\theta_{ijk}|D,S^h)$ (probabilidade condicionais das TPCs dos nós da estrutura da RB S^h).

O cálculo de $P(\theta_{ijk}|D,S^h)$ pode ser realizado conforme descrito na subseção anterior. O cálculo de $P(S^h|D)$ pode ser feito, via Regra de Bayes, da seguinte maneira:

$$P(S^h|D) = \frac{P(S^h)P(D|S^h)}{P(D)}$$ Equação 3.11

Sendo que $P(D)$ é independente da estrutura S^h. Dessa forma, para determinar a distribuição de probabilidade condicional das possíveis estruturas, é necessário calcular a verossimilhança marginal dos dados ($P(D|S^h)$) para cada estrutura (considerando a probabilidade a priori $P(S^h)$ de cada uma das estruturas) e então determinar a estrutura que codifica a DPC para $\boldsymbol{X}$ baseado nos valores máximos de verossimilhança. Este método, conhecido como abordagem Bayesiana completa (por considerar todas as possíveis estruturas de RB do espaço de hipóteses), é freqüentemente inviável (Heckerman, 1997). Por exemplo, para se ter uma idéia, o espaço de hipóteses das estruturas das RBs para um conjunto formado por apenas três variáveis *A*, *B* e *C* é formado por 25 diferentes estruturas. Na figura 3.11 são mostradas algumas dessa estruturas.

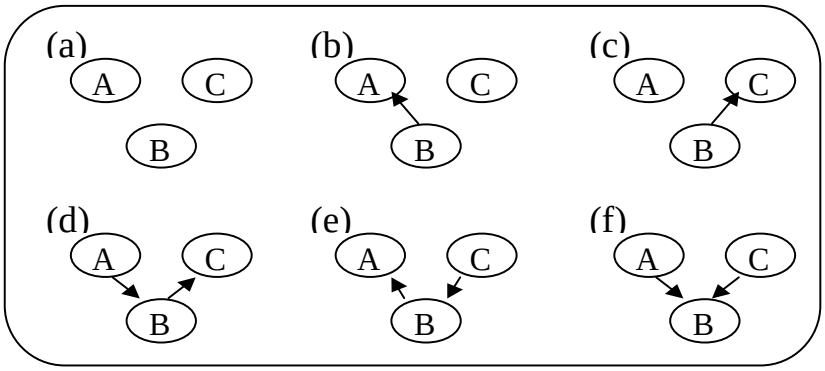

Figura 3.11. Algumas estruturas possíveis de redes Bayesianas para as variáveis *A*, *B* e *C*.

Em problemas com um grande número de variáveis a serem consideradas é necessário o uso de outras abordagens. Entre as mais referenciadas na literatura e mais utilizadas pelos métodos de aprendizado de RBs, destaca-se a abordagem de Seleção do Modelo.

A tarefa básica da Seleção do Modelo é selecionar um modelo "bom" (i.e. uma "boa" hipótese de estrutura) a partir dos modelos possíveis e considerá-lo como se fosse o modelo "correto". Esta abordagem pode suscitar várias questões:

- É possível obter resultados precisos na seleção do modelo com essa abordagem?
- Como decidir se o modelo é "bom" ou não?

- Como encontrar modelos “bons”?

A questão da precisão é difícil de ser respondida. No entanto, várias pesquisas mostram experimentalmente que a seleção de um único modelo “bom” freqüentemente produz resultados bastante satisfatórios (Heckerman et al., 1995).

Quanto à segunda questão, é necessário estabelecer algum critério para medir o grau em que uma determinada estrutura de RB combina adequadamente o conhecimento de fundo (estrutura da rede fornecida pelo especialista) com o conhecimento embutido nos dados. Um modelo bastante utilizado é o da probabilidade condicional relativa, calculada do seguinte modo:

Equação 3.12

Como pode ser observado na equação 3.12, esse critério possui dois componentes: o *log a priori* (*log* $P(S^h)$) e o *log de verossimilhança marginal* (*log* $P(D|S^h)$).

Vários métodos podem ser encontrados na literatura para o cálculo de *log* $P(S^h)$. O caso mais simples considera qualquer hipótese de S^h igualmente comum (não considerando, portanto, o valor de *log* $P(S^h)$ para o cálculo de *log* $P(D,S^h)$). Um outro método, proposto por (Heckerman et al., 1995), usa uma estrutura de rede dada a priori (conhecimento de fundo). A idéia básica deste método é penalizar a probabilidade ($P(S^h)$) de cada estrutura de acordo com alguma medida de desvio entre uma determinada estrutura e a rede considerada a priori. A métrica MDL (Descrição de Comprimento Mínimo), proposta por Lam e Bacchus (1994), faz balanceamento entre o ajuste aos dados e a complexidade do modelo. A adição de uma variável *pai* causa aumento do *log de verossimilhança*, bem como da penalização. Haverá adição de um arco caso o incremento na verossimilhança seja relevante.

Segundo Dawid (1984), o cálculo do *log* $P(D|S^h)$ poderia ser feito utilizando a equação 3.13:

Equação 3.13

O termo $P(x_e|x_1,...,x_{e-1},S^h)$ é a predição para x_e feita para o modelo S^h. O modelo com maior *log de verossimilhança marginal* (*log* $P(D|S^h)$), considerando as probabilidades *a*

priori das estruturas iguais, representa um "bom" modelo de estrutura para o conjunto de dados *D*.

Finalmente, com relação à terceira pergunta, várias pesquisas têm sido direcionadas para o uso de algoritmos de busca heurística, a fim de encontrar uma boa estrutura a partir do espaço de hipótese de todas as possíveis estruturas de RBs. Um exemplo desses algoritmos é o de *busca gulosa*, que inicia com a escolha de uma estrutura qualquer de RB. Então, avalia o *log de verossimilhança marginal* (através, por exemplo, da equação 3.13) para todas as possíveis mudanças *m* (e.g. adicionar ou retirar um arco da estrutura). Em seguida, realiza a mudança *m* cujo valor do *log de verossimilhança marginal* seja máximo. A busca é concluída quando não existe *m* que proporcione um valor maior para o *log de verossimilhança marginal*. Este algoritmo é utilizado por vários métodos de aprendizado de RBs como, por exemplo, o K2 (Cooper e Herskovits, 1992) e Bayesian Knowledge Discovery (Ramoni e Sebastiani, 1997).

Para ilustrar o aprendizado da estrutura de redes Bayesianas, considere o exemplo do K2. Este método aprende a estrutura de uma RB a partir de um conjunto de dados completo, cuja ordem das variáveis deve ser fornecida pelo usuário desse método. O K2 adota um método para calcular (selecionar) a estrutura com máxima verossimilhança e um algoritmo de aprendizado para encontrá-la (aproximadamente). Essa busca é iterativa e inicia com uma estrutura de rede bem simples, onde todas as variáveis são independentes uma das outras. Em seguida são avaliadas as verossimilhanças marginais de cada rede resultante de possíveis mudanças (e.g. adicionar uma nova ligação entre dois nós dessa rede), e então é aplicada a melhor das alterações antes de uma nova iteração. Esse processo continua até que o algoritmo não consiga encontrar nenhuma alteração simples que melhore (aumente) o valor de máxima verossimilhança.

Esse algoritmo, classificado como de busca e pontuação e um dos mais utilizados, permite encontrar a mais provável estrutura de rede de crença *S* a partir de um determinado conjunto de dados *D*. O algoritmo K2 aplica a pontuação Bayesiana segundo a equação 3.14.

Equação 3.14

na qual:

n é o número de nós;

q_i é o número de configurações dos pais da variável X_i;

r_i é o número de possíveis valores do nó X_i;

N_{ijk} é o número de casos em D onde o atributo X_i é instanciado com o seu valor k, e a configuração dos pais de X_i é instanciada com o valor j;

N_{ij} denota o número de observações em que a configuração dos pais de X_i é instanciada com o valor j, sendo $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

Após a construção da rede (a partir do conhecimento fundo do domínio, dos dados ou da combinação de ambos), são necessários mecanismos de inferência para computar as eventuais probabilidades de interesse. Na próxima seção, são apresentados alguns tipos de inferência em RBs.

3.3.6. Inferência em Redes Bayesianas

A tarefa básica de um sistema de aprendizado probabilístico é calcular a distribuição de probabilidade condicional de um conjunto de variáveis de uma RB, chamadas variáveis de consulta, dado os estados de um outro conjunto de variáveis, denominadas variáveis de evidência, que compõe esse modelo (Russell e Norvig, 2003). Por exemplo, no problema da detecção de fraude (figura 3.9), poder-se-ia obter a probabilidade de ocorrer determinada fraude, dada às evidências de compra de gasolina e o sexo do proprietário do cartão ser masculino. Como essa probabilidade não está diretamente armazenada na RB, é necessário computá-la. O cálculo de probabilidades de interesse em uma RB é geralmente conhecido como inferência probabilística.

A princípio, uma RB pode ser usada para calcular a distribuição probabilística para qualquer subconjunto de variáveis, dado os valores ou distribuição de qualquer subconjunto das variáveis restantes. A razão disso, conforme já foi apresentado na seção 3.5, é que as RBs determinam a distribuição de probabilidade conjunta de todas as variáveis representadas por ela.

Para melhor compreender o processo de inferência, considere a RB apresentada na figura 3.12, na qual todos os seus nós são binários.

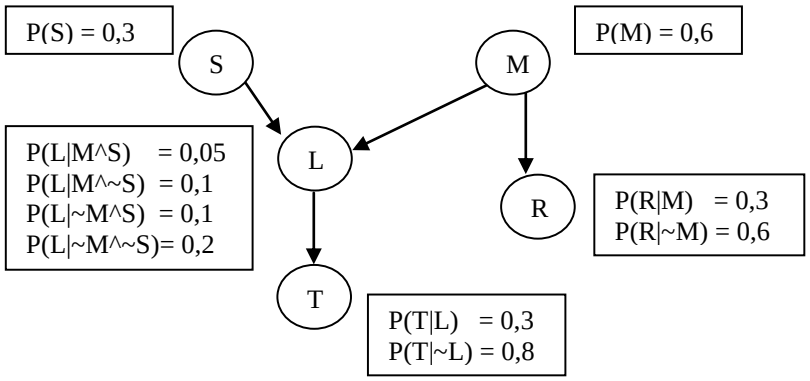

Figura 3.12. Rede Bayesiana utilizada como exemplo de um processo de inferência.

Caso fosse necessário realizar a seguinte inferência: "qual a probabilidade de *R* acontecer, dado que *T* ocorreu e *S* não ocorreu", ou seja, seria preciso calcular *P(R|T^~S)*. Esse cálculo poderia ser realizado, a partir das formulações do Teorema de Bayes, do seguinte modo:

Equação 3.15

É possível visualizarmos, a partir da equação abaixo, a situação em que A possua dois valores (estados) possíveis

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

Equação 3.16

A qual pode ser generalizada para os casos em que *A* tem *n* estados possíveis, conforme apresentado na equação 3.17.

Equação 3.17

Para o cálculo de *P(R|T^~S)*, é necessário computar as probabilidades a partir da distribuição de probabilidade conjunta das variáveis que compõem esta inferência, considerando a independência condicional entre os nós da RB.

Assim, aplicando a equação 3.15 para a variável binária *R*, obtém-se:

Equação 3.18

e considerando a independência condicional supracitada, o cálculo *P(R^T^~S)*, pode ser obtido do seguinte modo:

P(R^T^~S) = P(R^T^~S^L^M) + P(R^T^~S^L^~M) + P(R^T^~S^~L^M) + P(R^T^~S^~L^~M)

P(R^T^~S) = P(R|M)P(T|L)P(L|~S^M)P(~S)P(M)
+ P(R|~M)P(T|L)P(L|~S^~M)P(~S)P(~M)
+ P(R|M)P(T|~L)P(~L|~S^~M)P(~S)P(~M)
+ P(R|M)P(T|~L)P(~L|~S^~M)P(~S)P(~M)

Portanto, o cálculo de *P(R^T^~S)* poderia ser obtido utilizando-se os valores contidos nas tabelas de probabilidades condicionais da RB da figura 3.12. Analogamente, os outros termos da equação 3.18, poderiam ser calculados e, conseqüentemente, a inferência *P(R|T^~S)* ser computada.

O método aplicado nesse exemplo é conhecido como exato. Um método é denominado exato se realiza o cálculo das probabilidades condicionais por meio de somatórios e combinações de valores, sem outro erro associado, que não seja de arredondamento no cálculo (Castillo et al., 1997).

Vários métodos exatos de inferência têm sido propostos na literatura para RBs. Um dos métodos exato mais amplamente utilizados, inclusive em produtos de software, foi desenvolvido por Andersen et. al. (1989). Esse método aproveita a estrutura da rede para propagar as evidências, calculando probabilidades locais e evitando expressões globais, o que diminui sobremaneira o número de variáveis a serem manipuladas. Para isso, é necessário transformar a RB em uma estrutura de nós (chamada de árvore de junção), com características especiais, cujos nós são determinados por subconjuntos das variáveis da RB original. A estrutura da árvore de junção associada à rede original é fixa, sendo os cálculos realizados localmente no sentido de que um nó necessite se comunicar somente com os seus vizinhos.

Existe, entretanto, outras duas classes de métodos de inferência – métodos aproximados e métodos simbólicos. A escolha da classe do método depende fundamentalmente do domínio de aplicação, bem como considera, para efeito do

levantamento da complexidade e, em conseqüência, do custo computacional, os seguintes fatores (Nicholson e Jitnah, 1998):

- a estrutura da rede, isto é o número de nós e arcos;
- a cardinalidade das variáveis (número de estados das variáveis);
- o número de *clique tables(CT)*, isto é, a soma do número de entradas das tabelas de probabilidades condicionais associadas a todos os nós da RB. O número de *CT* de um nó *N* é dado por:

$$CT_N = NE \times PE_1 \times PE_2 \ldots PE_m \qquad \text{Equação 3.19}$$

Na qual, *NE* é o número de estados da variável *N* e PE_1, PE_2 .... PE_m é o número de estados de cada uma das variáveis cujo nós são pai de *N*.

Utiliza-se, ainda, como critério de medida de complexidade, o valor máximo de *CT*, representado por *MAX(CT)* de uma RB, isto é, o nó que apresenta o maior valor de *CT* na RB.

Para entendimento dessas métricas de complexidade das RBs, considere a RB, mostrada na figura 3.13, representada por 5 nós, identificados pelas variáveis binárias *A, B, C, D* e pela variável *E*, que possui 4 estados.

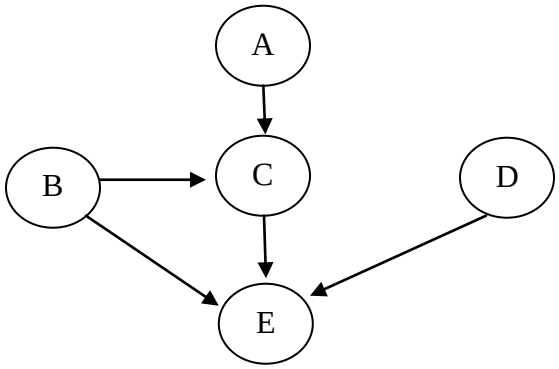

Figura 3.13. Rede bayesiana utilizada para instanciar as métricas de complexidade.

Pode-se observar que, na RB da figura 3.13, por exemplo, a variável *C* tem cardinalidade 2 e a variável E possui cardinalidade 4. Além disso, é possível dizer que o número de *CT* do nó *C* é 8 (2 x 2 x 2) e do nó *E* é igual a 32 (2 x 2 x 2 x 4) .

As demais medidas de complexidade da RB da figura 3.13 são as que são elencadas na tabela 3.3.

Tabela 3.3. Medidas de Complexidade da rede bayesiana da Figura 3.13.

Número de nós	Número de arcos	*MAX(CT)*
5	5	32

Na literatura, são encontrados uma série de estudos que tratam do uso de métodos de inferência sobre RBs com alta complexidade, nos quais, em sua maioria, são utilizados como *benchmark* algumas redes para essa finalidade e cuja complexidade é medida como na figura 3.3 (D´Ambrosio, 1999); (Dechter e Rish, 2003); (Ramos e Cozman, 2005). Na figura 3.4, são apresentadas algumas dessas RBs e seus parâmetros de complexidade.

Tabela 3.4. Medidas de Complexidade de um conjunto de RBs utilizadas como *benchmark*.

Rede Bayesiana	Número de nós	Número de arcos	*MAX(CT)*
Asia (Lauritzen e Spiegelhalte, 1988)	8	8	8
Alarm (Beinlich et al., 1989)	37	46	144
Hailfinder (Abramson et al., 1996)	56	66	3.267
Barley (Kristensen e Rasmussen, 1993)	48	84	13.063.680
Munin (Andreassen et al., 1989)	189	282	78.400.000

Outro componente que exerce influência substancial na complexidade computacional do método de inferência é a cardinalidade das variáveis. Por exemplo, a complexidade do método de árvore de junção é dado por $O(ce^m)$, no qual *e* é o número de estados de um nó da RB, *c* é o número de *CT* desse nó e *m* é o número máximo de *CT* da RB (Neapolitan, 90). Caso fosse aumentada a cardinalidade das variáveis da RB, o tamanho das TPCs da RB cresceria exponencialmente e, por conseguinte, o algoritmo de árvore de junção teria seu esforço computacional elevado de modo exponencial (Liu e Wellman, 2002).

Os métodos exatos de inferência em RB são classificados como NP-difícil. Portanto, para RBs com alta complexidade, é fundamental considerar a utilização de métodos aproximados de inferência, face à intratabilidade de redes extensas (alta complexidade), a partir do emprego de métodos exatos (Russell e Norvig, 2003). Embora os métodos de

inferência aproximados sejam também NP-difícil (Dagum e Luby, 1993), oferecem a oportunidade de considerar os problemas representados em RBs bem mais complexas, o que pode compensar substancialmente a perda de precisão.

Os métodos aproximados de inferência em uma RB sacrificam a precisão para poderem ganhar em eficiência. Por exemplo, o método Monte Carlo estabelece uma solução aproximada para amostragem aleatória da distribuição de variáveis não observadas. Os métodos aproximados podem ser classificados em algoritmos de simulação estocástica, métodos de simplificação de modelos, métodos baseados em busca e propagação de crença em ciclos (Castillo et al., 1997). Na literatura, diversos métodos são empregados para realizar inferência aproximada em RBs, com destaque para *likelihood weighting* (Fung e Chang, 1989), *Gibbs* Sampling (Russel e Norvig, 2003) e AIS Sampling (Cheng e Druzdzel, 2000).

Os métodos simbólicos, ao contrário dos métodos exatos e aproximados, são capazes de manipular parâmetros não-numéricos (simbólicos). Os métodos de propagação simbólica são particularmente recomendados quando não se tem bem definida a especificação numérica dos parâmetros do modelo probabilístico ou quando os especialistas somente são capazes de especificar intervalos dos parâmetros, ao invés de valores exatos.

Outros aspectos relacionados à inferência em RBs são também considerados na literatura, tais como a possibilidade do estabelecimento de correlações, considerando o tempo (Kjaeruff, 1992); (Santana et al., 2007) e a combinação com outros métodos com vistas à melhoria da interpretabilidade dessas redes, tais como sistemas fuzzy (Yang, 1997).

3.3.7. Redes Bayesianas e Mineração de Dados

Uma RB representa um modelo probabilístico completo das variáveis de um determinado domínio, podendo representar tanto informações qualitativas (dependências), quanto quantitativas (função de distribuição de probabilidades condicionais). Em adição, possui um mecanismo poderoso para realizar inferência sobre esse modelo. Essas características permitem com que as RBs possam ser utilizadas em MD não apenas como técnicas de raciocínio em situações que apresentam incerteza, mas também como uma espécie de estrutura de memória para o conhecimento extraído.

Outra característica relevante das RBs, é que elas consideram todos os atributos potencialmente relevantes, isto é, análises (inferências) não são associadas a classes absolutamente (a um único atributo).

Além disso, outros três fatores, descritos a seguir, têm motivado o uso de RBs em processos de MD (Heckerman, 1997).

Primeiro, a manipulação efetiva de conjuntos de dados incompletos. Por exemplo, o problema da classificação, utilizando técnicas de aprendizado supervisionado padrão, em um contexto em que duas variáveis de um conjunto de dados são fortemente correlacionadas, é particularmente simples quando todos os valores dessas variáveis são observados nesse conjunto. Entretanto, quando uma dessas variáveis não é observada, essas técnicas podem perder muito da precisão na classificação de novos exemplos, pois não codificam a correlação entre as variáveis. As RBs oferecem um modo natural de codificação dessa correlação (por meio de dependências entre variáveis de um conjunto de dados).

Segundo, o aprendizado sobre relacionamentos causais entre variáveis do domínio. Essa forma de aprendizado facilita o entendimento do domínio (e.g. em análise exploratória de dados), bem como permite efetuar previsões na presença de intervenções de especialistas de um domínio de aplicação (e.g. um analista de marketing pode desejar saber se deve ou não aumentar a exposição de um anúncio para elevar a venda de um determinado produto).

Terceiro, as RBs, em conjunto com outras técnicas de estatísticas Bayesianas, facilitam a combinação do conhecimento do domínio com os dados. As RBs possuem uma semântica causal que faz com que a codificação do conhecimento de fundo (causal) seja realizada de maneira direta.

Em especial, duas características foram determinantes para escolha dessa técnica para emprego nas investigações tratadas nesta tese, a saber: o auxílio ao processo de tomada de decisões baseadas em probabilidades da rede, quantificando, em termos probabilísticos, os efeitos de determinados eventos e a possibilidade de realizar análises sensitivas para entender qual aspecto do modelo tem maior impacto nas probabilidades das variáveis de consulta

3.4. CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentadas as técnicas para a resolução de problemas que envolvem a manipulação de incerteza, as quais foram empregadas nos processos de MD desenvolvidos nesta tese

A seguir, são apresentadas alguns estudos correlatos aos assuntos abordados neste trabalho, principalmente no que concerne ao aperfeiçoamento dos mecanismos de inferência de RBs e as aplicações dessa técnica no domínio de interesse desta tese – Sistemas Elétricos de Potência.

A idéia básica é apresentar o estado da arte das aplicações das técnicas empregadas nessa área, em especial as RBs e os algoritmos genéticos; e suscitar a discussão de outros temas relacionados à tese.

4. TRABALHOS RELACIONADOS

4.1. CONSIDERAÇÕES INICIAIS

Neste capítulo são apresentadas algumas referências que podem servir de base de comparação para as investigações propostas neste trabalho. São destacadas as contribuições desses trabalhos e quais os diferenciais das melhorias propostas no processo de inferências de redes bayesianas.

Além disso, são sublinhadas as aplicações das técnicas de modelagem de dependência e evolucionárias direcionadas à solução de problemas do setor elétrico, em suas várias categorias: geração, transmissão, distribuição e comercialização.

Os estudos correlatos estão organizados em duas seções, a saber: a primeira, que versa sobre os mecanismos utilizados para melhoria do processo de inferência nessas redes; e a segunda que focaliza as aplicações de técnicas de modelagem de dependências e evolucionárias para resolução de problemas inatos ao setor elétrico;

4.2. MECANISMOS PARA MELHORIA DO PROCESSO DE INFERÊNCIA EM REDES BAYESIANAS

São destacados nessa seção, alguns exemplos do uso de mecanismos para melhoria da compreensibilidade dos padrões descobertos, utilizando algumas técnicas de manipulação de incerteza combinadas com RBs.

É importante frisar que o emprego de técnicas evolucionárias, visando aumentar as potencialidades de inferência das RBs, não é muito difundido. Em grande parte dos trabalhos disponíveis, são investigadas aplicações dessas técnicas, em especial os algoritmos genéticos, para melhoria do processo de aprendizado da RB. Isto pode ser abalizado em (Handa e Katai, 2003), (Morales et al., 2004) e (Reiz et al., 2008).

Alguns trabalhos, entretanto, apontam para uma abordagem híbrida de métodos de inteligência computacional para otimizar e melhorar o processo de inferência. Por exemplo, Yang (1997) utiliza um mecanismo Bayesiano-Fuzzy para manipular valores de evidência contínuos nos processos de inferência.

No que concerne à descoberta da configuração mais provável das variáveis de uma RB, dada uma variável meta, podem ser destacados alguns trabalhos.

Apesar de não utilizar uma técnica de otimização combinada à inferência de RBs, um interessante modo de realizar essas inferências foi proposto em (Andersen et al. 1989). O método empregado implementa um mecanismo para identificar a configuração mais provável dos valores das variáveis de uma RB, dada uma ou mais variáveis de consulta.

Em (Abdelbar e Hedetniemi, 1997) é proposto um método de otimização híbrido, que combina AG com *Simulated Annealing*, para encontrar a configuração de estados das variáveis que melhor explicam a obtenção de um valor meta da variável de consulta. O artigo centraliza a discussão no desempenho do método, em detrimento do ganho de interpretabilidade da RB. Não oferecendo, com isso, nenhuma alternativa para o usuário de visualização dos cenários descobertos, além da combinação ótima dos estados das variáveis de evidência associada ao valor meta.

Em (de Campos et al., 2002), é abordada a descoberta das variáveis que mais influenciam na obtenção de um valor meta. Neste estudo, ao invés da mais provável configuração de todas as variáveis de evidência, o método, baseado em AG, é endereçado para encontrar um subconjunto das variáveis de evidências que proporcionam a máxima probabilidade de se obter um valor meta.

Já (Kuo et al., 2008) propõe uma solução, para o problema da busca da mais provável configuração das variáveis de evidência, por meio de um método de otimização baseado em medidas de entropia.

Outro detalhe importante, quando se trata das investigações sobre a busca pela configuração mais provável, está relacionado à complexidade. Em (Shimony e Domshlak, 2003), corroborado por (Dagum e Luby, 1993), os problemas são classificados como NP-difícil.

Muitos outros trabalhos podem ser citados na literatura, tais como (Mengshoel et al., 2006), (Verma e Rao, 2006) e (Pavón et al., 2009).

As principais diferenças em relação aos métodos encontrados na literatura e apresentados anteriormente e o método objeto de estudo desta tese, é que os primeiros miram-se no desempenho, por exemplo, (Abdelbar e Hedetniemi, 1997), (Shimony e Domshlak, 2003) e (Kuo et al., 2008), simplesmente na obtenção da configuração mais provável, casos de (Andersen et al. 1989), (Shimony e Domshlak, 2003), (Kuo et al., 2008) e (Pavón et al., 2009) ou na obtenção parcial da configuração mais provável (variáveis que exercem maior impacto sobre o valor meta), conforme as investigações de (de Campos et al., 2002). Já a abordagem desenvolvida aqui procura estender o poder de interpretabilidade da rede, oferecendo ao usuário diversos mecanismos de análise, tais como:

- a facilidade de uso de métodos de inferência em redes bayesianas exatos e aproximados como função de avaliação dos cenários;
- a possibilidade de encontrar as variáveis que exercem maior influência na obtenção de um valor meta;
- a capacidade de se obter os valores contínuos das variáveis de evidência;
- a incorporação de conhecimento do especialista para conduzir o processo de busca da melhor configuração das variáveis de consulta;
- a descoberta de cenários (configuração mais provável) com mais de uma variável meta, considerando os pesos, também estabelecido pelo especialista, de cada uma delas no contexto do domínio de aplicação.

Esses mecanismos podem ser vistos como um *framework* para estender o poder de interpretabilidade das redes bayesianas, fundamentalmente relacionado ao processo de busca de cenários ótimos para a obtenção de um valor meta para uma ou mais variáveis.

4.3. EMPREGO DE TÉCNICAS DE MODELAGEM DE DEPENDÊNCIAS E EVOLUCIONÁRIAS EM SISTEMAS ELÉTRICOS DE POTÊNCIA

Em razão das RBs oferecerem, dado o seu formalismo de representação de conhecimento, um mecanismo natural de modelagem de diagnósticos, a grande aplicação dessa técnica está associada à resolução dessa classe de problemas. No caso do domínio dos sistemas elétricos, há um emprego maciço em diagnósticos de falhas de equipamentos, além da

utilização desse método para prover análises qualitativas e quantitativas, para a avaliação da confiabilidade dos sistemas elétricos, com vistas às melhorias dos processos de planejamento e operação, conforme pode ser visto em (Yongli et al., 2008) e (Wilsona e Huzurbazar, 2007).

Nos estudos contidos em (Yongli et al., 2008), são utilizadas redes bayesianas com métodos aproximados de inferência, dada a complexidade das redes geradas, para avaliação de confiabilidade dos sistemas elétricos de potência. Dentre as principais vantagens do uso dessa abordagem, conforme aborda o autor, está relacionada ao cálculo do índice de frequência de falhas e no apoio a identificação dos componentes vulneráveis de um determinado sistema (linha de transmissão, transformador, disjuntor, gerador, etc).

Em (Yongli et al., 2006), é apresentada a aplicação de RBs para o diagnóstico de possíveis falhas de transmissão em sistemas elétricos. A principal motivação apresentada para o uso dessa abordagem é a facilidade com que os relacionamentos do tipo causa-efeito, principalmente em domínios com um elevado grau de incerteza, podem ser modelados.

Como forma de diminuir o tamanho das tabelas de probabilidades utilizadas no problema supracitado, um modelo de RBs é proposto com nós *Noisy-Or* e *Noisy-And*. Estes nós podem ser entendidos como uma generalização dos convencionais conectores lógicos *or* e *and*, respectivamente. A idéia é utiliza-los nas redes como elementos que podem simplificar a correlação entre as variáveis do sistema e a implicação das mesmas quanto ao surgimento de falhas de transmissão. Assim, ao invés de se fazer a relação direta de causa e efeito entre duas variáveis, por exemplo, as mesmas implicam para um nó *Noisy-Or* ou *Noisy-And* e essas conexões são parametrizadas, a partir do uso de probabilidades, quantificando o impacto que cada variável tem na possibilidade de ocorrência de falhas de transmissão.

Vale ressaltar que a definição da estrutura da RB e dos parâmetros iniciais depende do conhecimento a priori do especialista do domínio. Os parâmetros podem ser ajustados usando um algoritmo de aprendizado proposto no referido trabalho, similar ao algoritmo de treinamento de redes neurais artificiais *backpropagation.*

Em (Yonggiang et al., 2005), é apresentada uma aplicação de RBs também no contexto do diagnóstico de falhas, com ênfase nos possíveis defeitos que podem ocorrer no

funcionamento de uma classe importante de equipamentos elétricos – dos transformadores. Face à incerteza desse diagnóstico, gerada principalmente pela complexidade de configuração desses equipamentos, é necessário utilizar um método para auxiliar o especialista na análise da possibilidade de ocorrência de defeitos. No trabalho, é destacado que atualmente o método mais comumente utilizado é conhecido como DGA (*Dissolved Gas Analysis*). A idéia do método proposto por (Yonggiang et al., 2005) é criar um modelo de diagnóstico de falhas em transformadores, baseado em RBs e DGA. A construção da rede é realizada via interação com o especialista e os parâmetros são aprendidos a partir dos dados, considerando o conhecimento *a priori* decorrente dos principais motivos da ocorrência de falhas, tais como a alta ou baixa temperatura e descargas elétricas.

Várias outras aplicações de RBs em diagnóstico de falhas são investigadas na literatura, como os trabalhos mais recentes de (Flores-Loredo et al., 2005), (Flores-Quintanilla, 2005) e (Zhou et al., 2006).

Em (Zhou et al., 2006) as RBs são utilizadas não especificamente para o diagnóstico de falhas, mas para predizer a possibilidade de haver falha na distribuição de energia, considerando alguns aspectos climáticos. Para isso, é modelada uma RB de uma camada, conforme apresentada na figura 4.1, com o propósito de realizar predições de falhas (em 7 estados possíveis) a partir das condições de vento (em 4 estados) e da possibilidade de ocorrência de descargas atmosféricas (2 estados - sim ou não). São realizados também comparações com um outro modelo de previsão baseado em regressão com múltiplas variáveis.

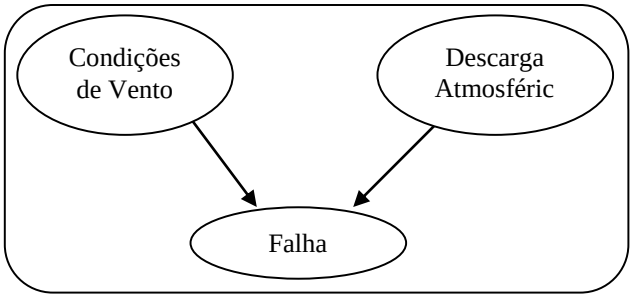

Figura 4.1. Rede Bayesiana de uma camada para predição de falha de distribuição de energia elétrica.

Os resultados obtidos pela RB foram similares ao da regressão, entretanto, são destacadas o maior poder de interpretabilidade das RBs, bem como a maior facilidade provida pelas mesmas para realizar ajustes no modelo (e.g. inclusão de novas variáveis).

No que concerne às soluções para a classe de problemas de otimização são considerados, via de regra, uma grande variedade de métodos exatos e técnicas heurísticas. No caso específico das técnicas heurísticas, é possível apresentar alguns exemplos de aplicações que envolvem técnicas evolucionárias de Inteligência Computacional.

Por exemplo, em (Mishra et al., 2007) são utilizadas técnicas evolucionários para minimizar as perdas de transmissão energia, com vistas a resolver uma classe de problemas conhecida como OPF (*Optimal Power Flow*). Para isso, são utilizados três tipos de técnicas: AG convencional, um AG modificado (utiliza a mutação como o operador dominante, ao invés do *crossover*) e BFA (*Bacteria Foraging Algorithm*), com o objetivo de otimizar a operação dos transformadores.

Em (Kumar et al., 2004), é utilizado um método que combina algoritmo genético com sistema fuzzy para pré-processar os dados recebidos por terminais remotos de medição de energia elétrica (medidores remotos), os quais realizam a aquisição dos dados e enviam para um servidor central. Em razão de falhas de transmissão ocorridas, em que quaisquer variações podem adicionar uma grande quantidade de ruído e ambigüidade nos dados adquiridos, é imperativa a realização do pré-processamento desses dados, antes de submetê-los aos demais sistemas de operação. Nesse método, o algoritmo genético é aplicado na saída do módulo fuzzy utilizado, a fim de que seja possível minimizar os erros ocorridos nas medições, antes de disponibiliza-las aos usuários.

Silva et al. (2006) propõe uma solução, usando algoritmo genético, para problemas de planejamento de redes de transmissão de energia. Esse problema consiste em encontrar um plano ótimo para a expansão de uma rede, considerando fatores como as linhas de transmissão, transformadores e demanda de energia, de tal modo que o sistema possa operar adequadamente. O algoritmo genético utilizado foi o proposto por Chu e Beasley (1997), pois, conforme os experimentos realizados, aderiu melhor ao problema supracitado. Esse modelo de algoritmo genético difere do padrão basicamente em função de três características:

- Usa uma função de aptidão (*fitness*) para identificar as soluções que aderem ao objetivo proposto e uma função de não-aptidão (*unfitness*) que quantifica as soluções não adequadas;
- Substitui, a cada geração, apenas um indivíduo de uma população; e
- Realiza uma estratégia para melhoria da aptidão de cada indivíduo avaliado.

Souza et al. (2006) destaca as potencialidades dos algoritmos genéticos na estimação de parâmetros relativo a sinais de tensão e corrente, de modo a atender às necessidades de energia elétrica dos consumidores da forma mais econômica possível, dentro de padrões compatíveis de segurança e qualidade. Os resultados obtidos comprovam que o algoritmo proposto pode identificar tais parâmetros com um alto grau de exatidão e eficácia para qualquer forma de onda de um sistema elétrico de potência, o que evidencia uma vantagem do algoritmo quando comparado a filtros dinâmicos (e.g. filtros de Kalman) que necessitam de reajustes em seus parâmetros.

Diversos outros trabalhos podem ser encontrados na linha de aplicação de algoritmos genéticos em sistemas elétricos destacada nessa seção, tais como os propostos por (Malachi e Singer, 2006), (Dahal et al., 2007) e (Yang et al., 2008), (Chiang, 2009).

As referências encontradas na literatura se caracterizam por oferecer um caráter exploratório para as análises, isto é, a partir do diagnóstico da situação (ou problema), toma-se uma decisão. Ao contrário, o método que desenvolvemos vislumbra a possibilidade de realizar análises antecipatórias. Assim, a partir de uma meta pré-estabelecida, podem-se identificar quais os cenários que propiciam obter tal meta, ou seja, quais os procedimentos devem ser adotados para alcançar os objetivos traçados antecipadamente.

4.4. CONSIDERAÇÕES FINAIS

Procurou-se abordar neste capítulo alguns trabalhos que discorrem sobre a aplicação de RBs e algoritmos genéticos no domínio do setor elétrico de potência, bem como algumas investigações que versam sobre extensões da interpretabilidade de RBs.

Entretanto, não estão muito bem definidas na literatura as soluções dos problemas das empresas deste setor que envolva a descoberta de quais fatores podem corroborar com a maximização de determinadas ações estratégicas das mesmas, dada à força da correlação existente entre essas ações e os demais aspectos que a influenciam, não obstante a imensa demanda, por parte dessas corporações, em obter soluções desta natureza.

Desse modo, no próximo capítulo, é apresentado um método híbrido que combina técnicas que possam estabelecer relações de causa e efeito (RBs) e de otimização (algoritmos genéticos), com vistas a encontrar os estados de determinadas variáveis que possam estabelecer uma condição ótima almejada para uma variável meta, influenciada por essas variáveis. No caso específico do setor elétrico, cuja aplicação será detalhada no Capítulo 6, para que seja possível atingir uma determinada meta de consumo de energia, determinar qual a configuração das variáveis que influenciam diretamente esse consumo.

5. ESTRATÉGIA PARA MELHORIA DA INTERPRETABILIDADE DE REDES BAYESIANAS

5.1. CONSIDERAÇÕES INICIAIS

As redes Bayesianas (RBs) têm sido amplamente utilizadas em domínios que envolvem incerteza e representa uma técnica consolidada na representação analítica do conhecimento obtido a partir dos dados, conforme as motivações para o seu uso em tarefas de Mineração de Dados, relacionadas na seção 3.3.7. Entretanto, as RBs possuem algumas restrições quanto ao modo em que são realizadas as inferências sobre elas.

Entre essas restrições, podem-se destacar a dificuldade de correlacionar as variáveis considerando o fator tempo e a dificuldade de estabelecer qual a combinação ótima de estados (cenário), para determinadas variáveis do domínio, que permita alcançar um determinado valor meta para uma (ou mais de uma) determinada variável do domínio.

Para atender a essas demandas, foi desenvolvido, no âmbito do Laboratório de Planejamento de Redes de Alto Desempenho (LPRAD), da UFPA, uma *framework* com o intuito de se estabelecer um conjunto de mecanismos de melhoria dos processos de aprendizado e de inferência em RBs, o que estamos tratando como estratégias de melhoria da interpretabilidade de RBs. Na figura 5.1, é apresentado esse *framework*, com seus respectivos componentes.

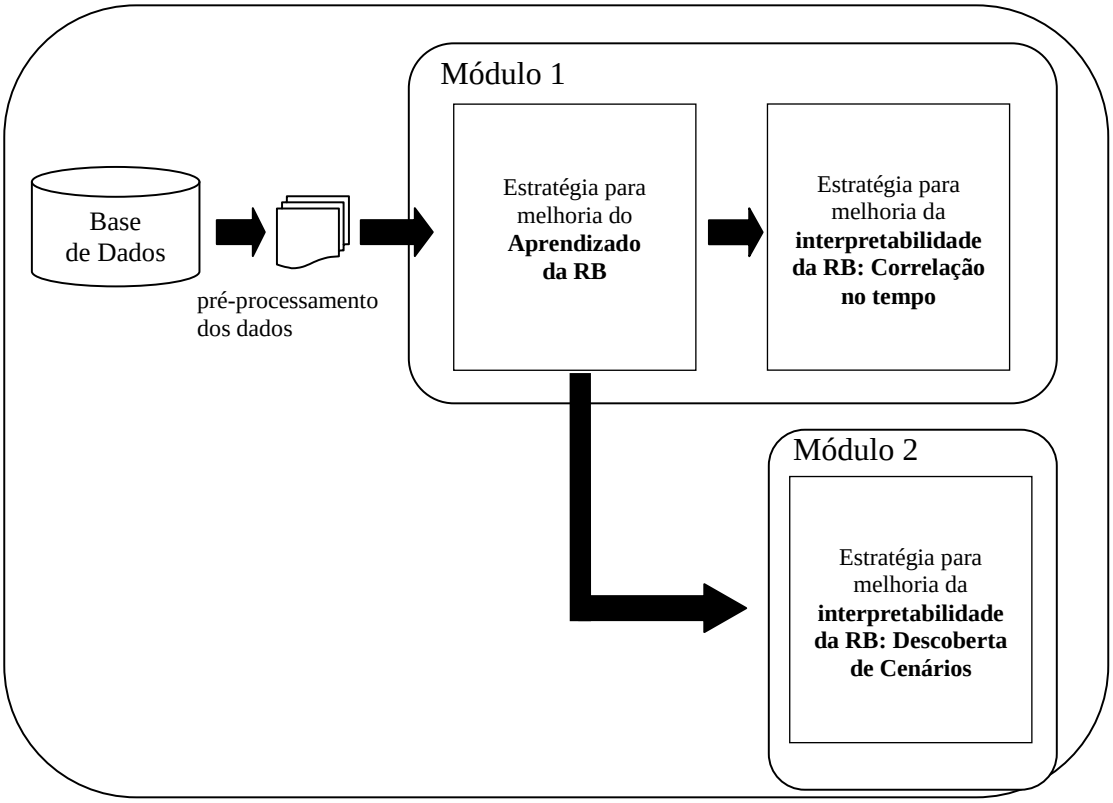

Figura 5.1. Enquadramento da estratégia desenvolvida no framework de aprendizado e inferência de RBs.

O módulo 1 desse *framework*, desenvolvida por (Santana, 2008), membro do grupo de Mineração de Dados do LPRAD, apresenta uma abordagem para estender o processo de inferência das RBs, a partir da consideração do tempo em que a correlação de eventos pode ocorrer. Em adição, o autor desenvolve um método para a melhoria do processo de construção (aprendizado) da rede a partir dos dados, por meio do algoritmo *Multiple Regression Structure Learner* (MRSL).

Neste capítulo, é apresentado uma estratégia original para estender o poder de inferência das RBs, relacionada à abordagem de otimização para a descoberta de cenários, a qual representa o módulo 2 do *framework* e é objeto desta tese de doutorado.

5.2. ABORDAGEM DE OTIMIZAÇÃO PARA A DESCOBERTA DE CENÁRIOS

A descoberta de cenários que propiciem a obtenção de uma determinada meta é de extrema importância para apoio ao processo de tomada de decisão. Por exemplo, determinar qual o conjunto de políticas públicas de saúde, educação e segurança necessárias à obtenção de uma meta de redução dos índices de violência de um estado é de grande valia para os gestores públicos, a fim de que possam tomar ações antecipadas e exitosas de combate à violência.

A abordagem desenvolvida aqui visa subsidiar os usuários de níveis decisórios quanto à possibilidade de, antecipadamente, poder analisar os cenários que podem acarretar a obtenção de um determinado objetivo. Para isso, foi proposto um método híbrido que combina o poder do relacionamento probabilístico provido pelas RBs, com a facilidade dos AGs para a incorporação do conhecimento específico do problema, com vistas à realização de tarefas de otimização.

As justificativas para o uso de algoritmos genéticos para realizar essa tarefa de otimização está calcada, principalmente, nas seguintes razões:

- São modulares e portáteis, no sentido que o mecanismo de evolução é separado da representação particular do problema considerado. Assim, eles podem ser transferidos de um problema para outro. Isto é particularmente interessante para o desenvolvimento desta tese em virtude dessa flexibilidade permitir a interação com diferentes tipos de mecanismos de inferência, conforme é visto na seção 5.2.2.

- São facilmente combinados com outras técnicas, incluindo as de inteligência computacional.

- A simplificação que o AG permite na formulação e solução de problemas de otimização, isto é, a facilidade da incorporação do conhecimento específico do problema nos procedimentos elaborados pelo AG (e.g. representação e avaliação das soluções), o que é importante quando se deseja especializar o método de otimização para o domínio a ser empregado, visto que uma solução de otimização genérica é teoricamente impossível (Wolpert e Macready, 1997).

- Os AG's possuem um paralelismo implícito, decorrente da avaliação independente de cada um dos indivíduos (soluções), ou seja, pode-se avaliar a viabilidade de um conjunto de parâmetros para a solução do problema de otimização em questão. O que é bastante recomendado para a descoberta da configuração ótima dos estados das variáveis de uma RB, principalmente se considerarmos os métodos aproximados de inferência, frequentemente associados às redes com um elevado número de variáveis e, consequentemente, espaços de soluções de dimensões elevadas.
- Os AG's são numericamente robustos, ou seja, não são sensíveis a erros de arredondamento no que se refere aos seus resultados finais.

Como utilizamos a estratégia desenvolvida em atividades de descoberta de padrões nos dados, mediante um processo de mineração, na próxima seção é descrita o enquadramento do mesmo no contexto desse processo.

5.2.1. Enquadramento da Estratégia de Descoberta de Cenário no Processo de Mineração de Dados

Pode-se dizer que a estratégia de descoberta de cenários de otimização aqui desenvolvida está relacionada às etapas Extração de Padrões e de Pós-Processamento do Processo de MD, apresentadas na seção 2.2. Quanto à etapa de Extração de Padrões, a abordagem permite a realização de inferências sobre RBs e a descoberta de cenários, gerados a partir dos padrões embutidos nos dados. No que concerne ao Pós-Processamento, decorre-se dessa estratégia a melhoria da compreensibilidade dos padrões gerados a partir dos dados, via RBs, na medida em que facilita a interpretação dos cenários por parte dos usuários, o que pode ser aferido por meio de medidas subjetivas de interessabilidade dos padrões descobertos.

É importante ressaltar que o Framework, apresentado na Figura 5.1 contempla especificamente as etapas de extração de padrões e pós-processamento, no que pese que todas as demais etapas devam ser contempladas (compreensão do domínio de aplicação, pré-processamento e utilização do conhecimento extraído) em um processo completo de uma aplicação de MD.

Feita a indicação do posicionamento da estratégia no contexto do processo de MD, é apresentado a seguir o arcabouço que constitui essa abordagem de descoberta de cenários.

5.2.2. Descrição da Estratégia de Descoberta de Cenários

A estratégia aqui desenvolvida busca identificar a melhor configuração, dentre os possíveis valores das variáveis existentes no domínio, que corroborem com a obtenção de um valor meta para uma (ou mais) variável (eis) restantes do domínio em questão.

Em razão do processo de descoberta dos cenários utilizarem os mecanismos de inferência da própria RB e com vistas a estabelecer uma abordagem o mais abrangente possível, quanto à classe de problemas que ele pode resolver, a estratégia permite o emprego tanto de métodos exatos, quanto de métodos aproximados de inferência.

Conforme é abordado na seção 3.3.6, existe uma grande variedade de algoritmos de inferências em RB na literatura, sendo a escolha dependente do domínio de aplicação e, em decorrência, da complexidade da RB que representa tal domínio.

Para efeito da estratégia de descoberta de cenários, é possível utilizar métodos de inferência exatos ou aproximados. Assim, o método de inferência recebe como entrada a solicitação de inferência $P(X|E)$, no qual X representa a variável meta e E o conjunto de variáveis de evidência, e devolve o valor correspondente a essa consulta. Esse valor é utilizado como função de aptidão dos indivíduos (cenários) do algoritmo genético. Na figura 5.2, é ilustrado um esquema de representação do processo de descoberta de cenários.

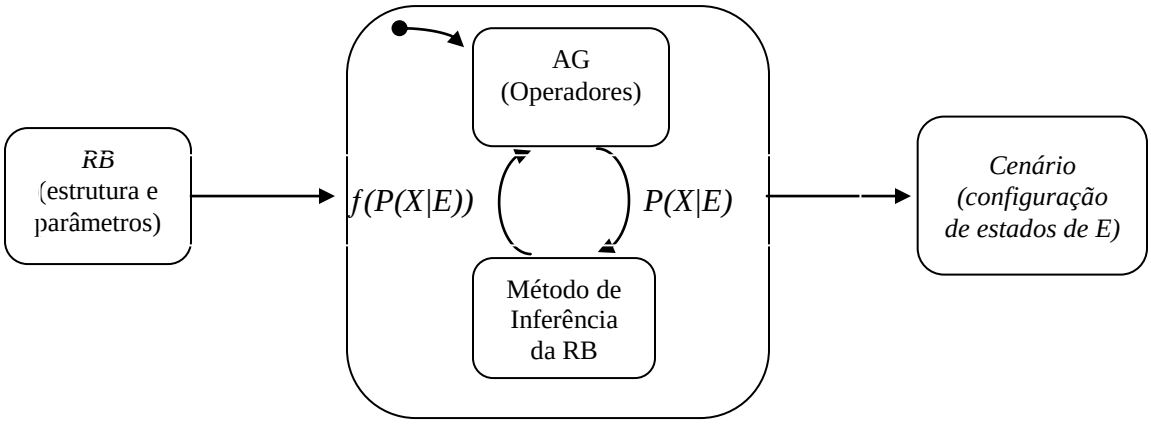

Figura 5.2. Esquema de representação da estratégia de descoberta de cenários.

Conforme pode ser observado na Figura 5.2, o processo de descoberta de cenários é iniciado com o fornecimento da RB, gerada a partir dos dados, e seus respectivos parâmetros, em seguida é aplicado um AG que, utilizando como função de aptidão dos cenários (indivíduos) o próprio mecanismo de inferência da RB, obtém, ao término das suas iterações, o cenário ótimo à obtenção de uma determinada meta.

O Método utilizado na estratégia desenvolvida aqui pode ser enquadrado na categoria dos métodos híbridos intercomunicativos, ou seja, os paradigmas (RB e AG) são usados em diferentes subsistemas, os quais trabalham em colaboração para chegarem a uma solução, ou seja, os paradigmas inteligentes são independentes e estes trocam informações e realizam funções separadas para gerar soluções, conforme ilustra a figura 5.2. Neste modelo, os paradigmas se comunicam com bastante intensidade e frequência. Como forma de mostrar essa interação entre os métodos que compõe a estratégia de descoberta de cenários, é apresentado, na Figura 5.3, um algoritmo que codifica o relacionamento entre o AG e o método de inferência da RB.

```
1.  DESCOBERTA_CENÁRIO (rb)
2.  /* retorna o cenário que melhor contribui para obtenção do valor meta para uma
    variável do domínio de aplicação */
3.  // argumentos: rb //rede bayesiana que codifica a distribuição conjunta P(X1, X2, ...Xn)
4.  população ← POPULAÇÃO_INICIAL_ALEATÓRIA;
5.  repita
6.    nova_população ← conjunto vazio
7.    para i ← 1 até TAMANHO(população) faça
8.      a ← SELEÇÃO(população, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(rb))
9.      b ← SELEÇÃO(população, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(rb))
10.     se (TAXA_CROSSOVER é atendida)
11.       filho_ab ← CROSSOVER(a,b)
12.     se (TAXA_MUTAÇÃO é atendida)
13.       filho_ab ← MUTAÇÃO(filho_ab)
14.     incluir filho_ab em nova_população
15.   fim para
16.   população ← nova_população;
17. até algum indivíduo (cenário) está adaptado o suficiente ou até ser obtido um
    número n de gerações
18. retornar o melhor cenário em população, de acordo com FUNÇÃO_APTIDÃO
```

Figura 5.3. Algoritmo do processo de descoberta de cenários.

Assim, o AG começa com a geração aleatória da população inicial *I*, constituída por um conjunto de cenários candidatos, os quais são, a seguir, avaliados, por meio do método de inferência da RB, a fim de obter a aptidão de todos os cenários, medida pela probabilidade de obtenção do valor meta da variável de consulta *X*, dada uma determinada configuração de estados (cenário) das variáveis de evidência *E*. O processo continua com a seleção dos indivíduos, por meio do método da roleta. Em seguida, são aplicados os operadores de *crossover*, com taxa de *crossover* T_c, e de mutação, com uma taxa de mutação T_m. O Processo se repete por *n* gerações e é finalizado seguindo um dos três critérios abaixo:

- estabelecimento de um número pré-determinado de gerações, ou seja, é definido um valor *a priori* para *n*;
- convergência do algoritmo, isto é, quando a aptidão de 90% dos indivíduos de uma população estiverem dentro de uma margem estabelecida pelo usuário.
- até que o algoritmo possa encontrar um cenário aceitável. A aceitação do cenário é realizada considerando um método subjetivo de avaliação de interessabilidade do padrão descoberto (cenário), o qual será explicado na seção 5.2.6.

Vale ressaltar que os parâmetros utilizados para a execução do AG são definidos pelo usuário, variam de acordo com o domínio de aplicação e, em última análise, a interessabilidade dos padrões descobertos (cenário).

Pode-se notar que é possível empregar um método qualquer de inferência em RB, representada em destaque no algoritmo da figura 5.3 por *MÉTODO_INFERÊNCIA_I*, quer seja exato ou aproximado, informando apenas o valor da probabilidade de ocorrência de um dado cenário para obtenção do valor meta de uma determinada variável de consulta, probabilidade essa que é utilizada como instrumento de cálculo da função de avaliação dos indivíduos do AG.

Com intuito de mostrar o processo genérico de interação do AG com o método de inferência da RB, considere uma rede bayesiana *B*, gerada a partir de um conjunto de dados *D*. Considere, ainda, o processo geral de inferência sobre *B*, expresso por um conjunto de

variáveis de consulta *X*, um conjunto de variáveis de evidência *E*, um conjunto *e* de valores observados para *E* e um conjunto *Y* de variáveis restantes (não contidas em *X* e *E*). Desse modo, uma consulta *P(X|e)* pode ser calculada por:

$$P(X \mid e) = \alpha P(X, e) = \alpha \sum_{y} P(X, e, y) \qquad \text{Equação 5.1}$$

Na qual α é uma constante de normalização, que assegura que a soma para a distribuição de probabilidade para *P(X|e)* seja igual a 1 e *y* são os valores possíveis para as variáveis representadas por *Y*.

A equação 5.1 pode realizar inferências pontuais sobre uma variável de consulta *X*, a partir de um conjunto qualquer de variáveis de evidência *E*, considerando para o cálculo da inferência as variáveis *Y*, conforme estabelece a equação 5.1. A estratégia de descoberta de cenários pode ser vista como uma especialização dessa equação, a qual se propõe a encontrar quais valores (estados) *e* do conjunto de variáveis *E* proporcionam a obtenção de um estado desejado de *X*. Nesse caso, *E* é formado por todas as variáveis do domínio, isto é $Y = \varnothing$, exceto as variáveis de consulta *X*, que representam justamente a meta a ser encontrada. Com isso, podemos escrever a equação 5.1, conforme abaixo (equação 5.2), considerando a avaliação da aptidão de um determinado indivíduo do AG que propicie a obtenção do valor meta x_i,

$$P(x_i \mid e_1, e_2, ..., e_n) = P(x_i) \prod_{k=1}^{n} P(e_k \mid x_i) \qquad \text{Equação 5.2}$$

na qual:

$e_1, e_2, ..., e_n$ são as possíveis evidências e;

x_i é o evento que queremos observar.

Com vistas a efetuar o processo de busca de cenários, é empregado um algoritmo genético, nos quais os cromossomos, representados por valores decimais, são constituídos pelos estados das variáveis de evidências, conforme a figura 5.4, na qual e_1 representa um estado qualquer de E_1, e_2 representa um determinado estado de E_2 e assim sucessivamente.

e_1	e_2	e_3	...	e_m

Figura 5.4. Representação dos cromossomos do algoritmo genético.

Para o cálculo da aptidão, os cromossomos são submetidos ao módulo de inferência da RB, com a finalidade de calcular a probabilidade da variável de consulta assumir determinado valor desejado (meta). Quanto maior essa probabilidade, mais apto será considerado esse indivíduo. Cabe ressaltar que podem ser utilizadas mais de uma variável de consulta, para efeito da descoberta de cenários, o que é demonstrado na seção 5.2.4.

Para ilustrar o funcionamento a estratégia de descoberta de cenários, considere a RB, apresentada na Figura 5.5(a). Em adição, observe os estados das variáveis representadas pelos nós da RB (Figura 5.5 (b)).

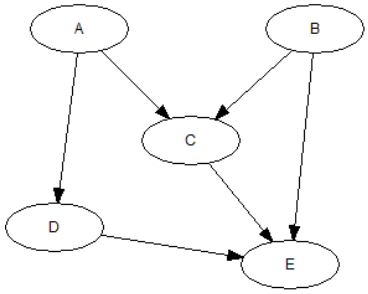

Figura 5.5(a). Rede bayesiana para exemplificar a descoberta de cenários

Variável	Estados
A	a_1, a_2
B	b_1, b_2, b_3, b_4
C	c_1, c_2
D	d_1, d_2
E	e_1, e_2, e_3, e_4

Figura 5.5(b). Estados das variáveis da RB

No exemplo, d_1 será considerado o valor meta, destacando-se que seria possível escolher qualquer (quaisquer) variável (eis) da RB. O AG atua sobre o método de inferência da RB (e.g. o método exato Árvore de Junção) para encontrar o cenário que propicia a obtenção de d_1, com a máxima probabilidade (caso ótimo).

Um cromossomo que representa uma possível solução (cenário candidato ao cenário ótimo) poderia ser representado pela figura 5.6. Na qual, a primeira posição (gene) define o estado a_2 da variável *A*, a segunda o estado b_3 da variável *B*, a terceira o estado c_1 para variável *C* e o quarto gene estabelece o estado e_2 para a variável *E*.

2	3	1	2

Figura 5.6. Representação de um possível cromossomo (cenário candidato) do AG.

A avaliação da aptidão, lembrando que a função de custo do AG é a própria RB e seu mecanismo de inferência, será dada por $P(d_1 | a_2,b_3,c_1,e_2)$. Assim, após aplicação dos operadores do AG (seleção, *crossover* e mutação) e ao término das iterações (gerações)

desse AG, obter-se-ia a melhor configuração (cenário) dos estados das variáveis *A, B, C* e *E*, que maximizasse a probabilidade de ocorrência de d_1.

O exemplo em questão, evidentemente, possui um espaço de busca bastante reduzido para o AG e foi utilizado unicamente para apresentar, de modo prático e didático, o seu funcionamento. Entretanto, em muitos domínios reais a complexidade envolve um espaço de busca consideravelmente maior. Por exemplo, a RB gerada para realização de diagnósticos preliminares de doenças dos músculos e nervos, desenvolvida por (Andreassen et al., 1989) e conhecida como Munin, possui um complexidade bastante elevada, conforme pode ser visto na tabela 3.4. Na RB Munin muitos dos seus nós têm até 20 estados, além de possuir um *MAX(CT)* bastante elevado (78.400.000), o que comprova um espaço de busca que justifica o uso de métodos de otimização, como os AGs.

Para comprovar a viabilidade da aplicação da estratégia em RBs que possuam alta complexidade, bem como do uso do AG combinado com métodos de inferência aproximados de RBs, foi utilizada a RB do Munin, amplamente utilizada como *benchmark* para avaliação desse tipo de inferência. Conforme pode ser visto na seção 3.3.6, a Munin apresenta uma das maiores complexidade, face ao seu número de nós, arcos e *MAX(CT)*. Na figura 5.7, é apresentada a arquitetura da RB Munin. Em virtude do número de nós e de arcos, a figura 5.7 se presta a ilustrar tal complexidade e não propriamente as variáveis que compõem os nós ou outra informação relativa aos parâmetros da RB (a estrutura da RB Munin é descrita em ANDREASSEN, 1989).

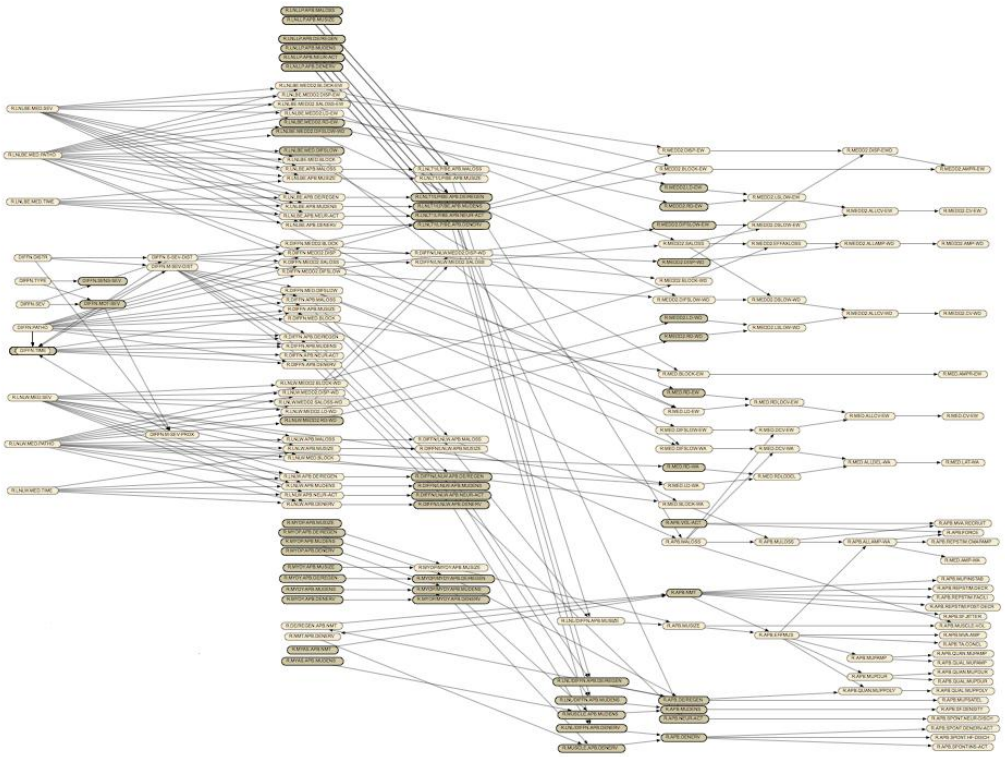

Figura 5.7. Rede bayesiana Munin.

Considere como exemplo a necessidade de se descobrir o cenário que melhor explica a ocorrência do tipo sensorial da neuropatia difusa[2], de caráter motor (*DIFFN_TYPE* = "MOTOR"). Após a execução da estratégia de descoberta de cenários, é possível identificar um dos cenários que possibilitam encontrar a máxima probabilidade de ocorrência de *DIFFN_TYPE* = "MOTOR", no caso foi obtido o valor 1 (100%), mostrado na figura 5.8.

```
1 3 1 1 4 1 2 M 4 1 3 1 2 2 2 3 2 1 1 2 4 5 1 1 2 2 3 4 3 4 5 4 2
1 1 3 3 4 1 1 1 3 2 3 1 3 1 3 1 2 1 1 1 2 1 2 5 4 1 2 2 2 2 2 1 3
4 3 4 2 1 2 1 3 2 5 3 2 3 2 5 2 1 3 3 1 4 5 2 2 1 1 1 1 6 3 5 4 2
3 3 3 1 1 4 2 3 1 2 4 2 4 5 1 3 2 2 1 3 2 2 1 2 4 3 4 4 3 5 3 1 1
3 3 3 2 7 2 1 1 3 3 1 2 5 3 1 3 2 5 5 3 7 7 3 2 31 27 2 2 3 1 3 3
7 5 4 6 2 1 2 4 1 3 3 6 27 5 7 2 7 3 4 5 6 3 29 5 9
```

Figura 5.8. Configuração das variáveis (cenário) que possibilita obter a máxima probabilidade da ocorrência de *DIFFN_TYPE* = "MOTOR".

É possível observar na figura 5.8 que cada valor representa um estado de uma variável da RB Munin e que *M* é a variável observada *DIFFN_TYPE* (meta). Assim, por

[2] Tipo de patologia que afeta os nervos periféricos, principalmente os dos pés e pernas, causando perda sensorial e/ou motora.

exemplo, o último valor da figura 5.8 identifica o estado 9 da variável que representa o último nó da RB Munin.

Para executar o AG sobre a RB Munin, foi empregado o método de inferência aproximado *AIS Sampling*, como função de aptidão dos indivíduos (i.e. *MÉTODO_INFERÊNCIA_I = AIS Sampling).* Vale destacar que o AG desenvolvido neste trabalho pode utilizar qualquer outro método de inferência (exato ou aproximado) como função de aptidão. Dada a complexidade da rede Munin, o AG foi implementado de acordo com os seguintes parâmetros (Tabela 5.1). Não esquecendo que apesar de utilizarmos nessa aplicação o algoritmo aproximado *AIS Sampling*, poder-se-ia utilizar qualquer outro método aproximado, como o algoritmo CMMC e de Ponderação de Probabilidade.

Tabela 5.1. Parâmetros utilizados no *AG* para a RB MUNIN.

Parâmetro	**Valores**
População inicial	500 indivíduos
Número de gerações	10000 gerações
Seleção	método da roleta
Crossover	cruzamento de um ponto
Taxa de *Crossover*	90%
Taxa de mutação	0,01%

Cabe acrescentar que a estratégia de otimização, objeto desta tese, elaborada para aperfeiçoar o processo de interpretabilidade de redes bayesianas também contempla quatro outras abordagens úteis para determinados domínios de aplicação e que permitem:

1. Identificar as variáveis de evidência que exercem maior impacto na obtenção de um valor meta. Essa abordagem, apresentada na seção 5.2.3, é particularmente útil em situações em que é estabelecida uma meta a ser atingida, mais nem todos os estados das variáveis de evidência são conhecidos ou gerenciáveis. Além de ser interessante em redes com muitos nós, caso da RB Munin utilizada como exemplo, cuja análise dos cenários descobertos com todas as variáveis, por parte dos usuários de níveis decisórios, fica bastante comprometida ou impraticável.

2. Encontrar o cenário, com valores contínuos, dentro das faixas (estados) de cada uma das variáveis de evidência, que mais contribuem para a obtenção de um

valor meta para a variável de consulta. Essa abordagem é utilizada em domínios que envolvam valores numéricos e é abordada na seção 5.2.4.

3. Estender a estratégia desenvolvida para a obtenção de valores meta para mais de uma variável de consulta. Essa abordagem, descrita na seção 5.2.5 é interessante quando se deseja estabelecer mais de uma meta, ponderando a importância de cada uma delas, de acordo com a avaliação do usuário.

4. Embutir o conhecimento especialista (conhecimento de fundo) no processo de descoberta de cenários, de tal modo que sejam utilizados critérios subjetivos para avaliar os cenários. Essa abordagem, explicada na seção 5.2.6., em última análise, estabelece um critério de parada pautado no grau de interesse, medido a partir da crença de algum aspecto do domínio relacionado à variável meta, estabelecida *a priori* pelo especialista.

A motivação maior para o desenvolvimento dessas abordagens está relacionada às demandas do setor elétrico, foco de aplicação principal desta tese, a qual será realizada no âmbito do Capítulo 6.

5.2.3. Identificação das variáveis de maior influência sobre a variável meta

Alternativamente, a estratégia de descoberta de cenários pode identificar o conjunto de variáveis de evidência M que exercem maior impacto sobre a variável meta, independente das demais variáveis do domínio. Desse modo, é realizada uma modificação na função de avaliação do AG, mais especificamente na forma como são realizadas as inferências. Ao invés de se efetuar as medidas de aptidão dos indivíduos por meio de $P(X|E,Y)$, com $Y = \varnothing$ (i.e. E é formado por todas as variáveis do domínio, exceto X), considera-se, para o espaço de busca, os indivíduos que propiciem a maior probabilidade de obtenção do valor alvo, independente do número de variáveis de evidências E.

A motivação basilar dessa abordagem é identificar, no conjunto de variáveis E, um subconjunto M das variáveis que exercem maior influência sobre X, independente das demais variáveis Y. Em outras palavras, é possível encontrar os estados das variáveis de E

que permitem encontrar a máxima probabilidade de ocorrência do estado meta de *X*, independente dos estados de *Y*, no qual $E = Y \cup M$.

Para realizar a avaliação dos indivíduos do AG, é utilizada a equação 5.1 e não a sua especialização, codificada na equação 5.2, pois essa considera todas as variáveis *E*. Com isso, o algoritmo DESCOBERTA_CENÁRIO (figura 5.3) passa a considerar os indivíduos com qualquer combinação de evidências ($P(X|E,Y)$). O procedimento utilizado para realizar essa alteração consiste do estabelecimento de um *flag* (*0* - zero) nas variáveis de evidência. Assim, por exemplo, a medida de aptidão de um indivíduo ($e_1,0,e_3,e_4$) é dada por $P(x_1|e_1,e_3,e_4)$, desconsiderando a variável e_2 no cálculo da inferência. De outro modo, podemos dizer que um novo estado foi atribuído às variáveis *E*, no caso representado por *0*.

Com o objetivo de exemplificar o uso da abordagem tratada nessa seção, considere a RB Asia (Lauritzen e Spiegelhalte, 1988), mostrada na figura 5.9, também largamente referenciada na literatura como *benchmark*, e a tarefa de encontrar as variáveis que mais influenciam na obtenção da máxima probabilidade de ocorrência de (raio-X= sim).

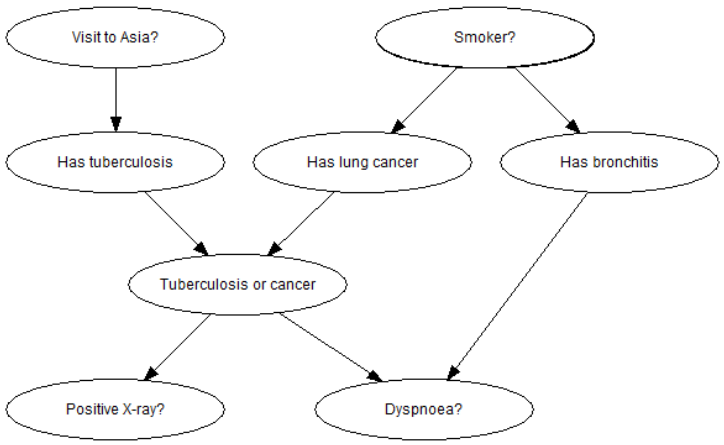

Figura 5.9. Rede Bayesiana Asia.

Considere também que após o término das iterações do AG (depois de *G* gerações) obtenham-se os indivíduos (por simplificação, são mostrados apenas 10 indivíduos da última geração do AG) da tabela 5.2.

Tabela 5.2. Indivíduos gerados após *G* gerações do AG.

Indivíduo								Aptidão
fumante	visitaAsia	tuberculose	câncer	tuber. oucâncer	raio-x	bronquite	dispneia	*P(X|E,Y)*
0	**1**	**2**	**1**	0	M	0	1	0,951456
1	**1**	**2**	**1**	1	M	1	2	0,951456
0	**1**	**2**	**1**	1	M	1	2	0,951456
1	**1**	**2**	**1**	0	M	1	1	0,951456
0	**1**	**2**	**1**	1	M	1	1	0,951456
1	**1**	**2**	**1**	1	M	1	1	0,951456
1	**1**	**2**	**1**	1	M	1	2	0,951456
1	**1**	**2**	**1**	1	M	1	1	0,951456
1	**1**	**2**	**1**	1	M	1	1	0,951456
1	**1**	**2**	**1**	1	M	1	2	0,951456

Observa-se, a partir da tabela 5.2, que o AG encontrou a mais provável configuração (1, 1, 2, 1, 1, 1, 1 ou 2) para obtenção de M (raio-X= sim). Além disso, pode-se notar (em destaque na tabela 5.2) que as variáveis v*isitaAsia*, *tuberculose* e c*âncer*, quando instanciadas com v*isitaAsia=1(não), tuberculose=2*(sim) e c*âncer=1(não),* são determinantes para definição da maior probabilidade de alcance do valor meta (raio-X= sim) – *0,951456*. Pode-se dizer, portanto, que as variáveis v*isitaAsia, tuberculose* e c*âncer,* quando combinadas, exercem maior impacto na meta a ser alcançada.

Outra utilidade observada na descoberta de cenários parciais está relacionada à sua aplicação em redes com muitos nós, na qual a análise do cenário completo é de difícil compreensão humana. Por exemplo, no caso da RB Munin, no qual um possível cenário completo é mostrado na figura 5.8, é possível identificar, por meio desta abordagem quais as variáveis que mais evidenciam a probabilidade máxima de ocorrência da meta *DIFFN_TYPE* = "MOTOR". Utilizando a mesma abordagem demonstrada no caso da RB Asia, é possível identificar as variáveis que exercem maior influência sobre a variável meta. Nas figuras 5.10, 5.11 e 5.12, é possível observar, em destaque, os três indivíduos com maior aptidão (máxima probabilidade para obtenção do estado meta *DIFFN_TYPE* = "MOTOR").

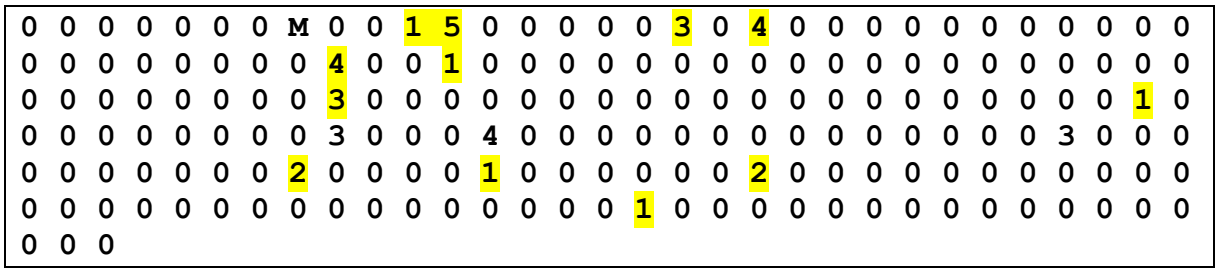

Figura 5.10. Indivíduo com aptidão igual a 0,976515.

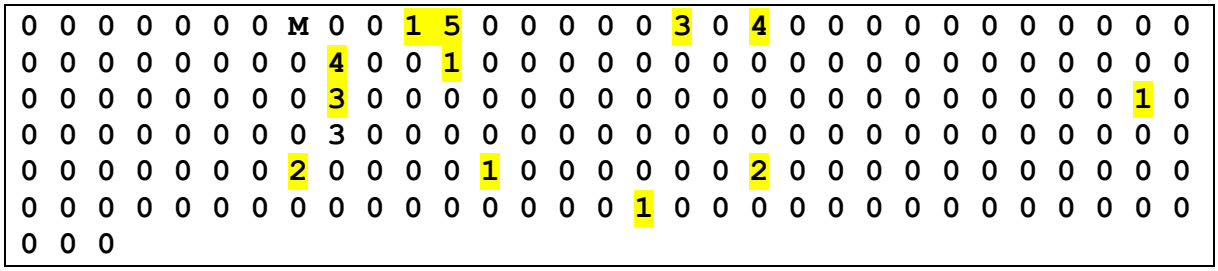

Figura 5.11. Indivíduo com aptidão igual a 0,96616.

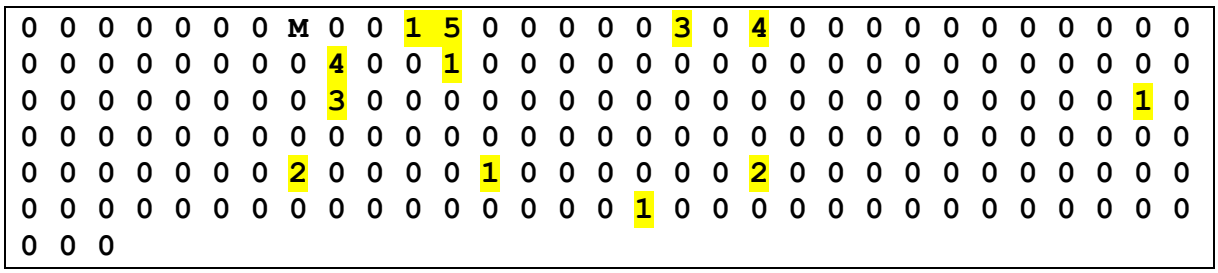

Figura 5.12. Indivíduo com aptidão igual a 0,989246.

Desse modo, deduz-se que as variáveis em destaque nas figuras 5.10, 5.11 e 5.12 podem ser consideradas as mais influentes, dada a alta probabilidade de obtenção do valor meta nas configurações de estados dos indivíduos, representadas pela combinação dos valores das variáveis em destaque nessas figuras.

5.2.4. Descoberta de Cenários – Valores Numéricos

Conforme pode ser observado na figura 5.3, o algoritmo básico da estratégia desenvolvida retorna como solução o cenário ótimo para obtenção de um estado desejado da variável de consulta (variável meta).

Entretanto, esse cenário é descrito pelos estados (discretos) das variáveis de evidência. Por exemplo, se fosse considerada a RB da figura 5.5 (a) e seus estados, representados na figura 5.5 (b), como sendo faixas de valores numéricos, ou seja, a_1 poderia

representar, por exemplo, o intervalo de valores reais [1, 25[e a_2 outro intervalo compreendido entre [25, 48]. Desse modo, ao invés de o método retornar um cenário ótimo com os valores discretos, por exemplo a_1, b_4, c_2 e e_1, ele retornaria esse cenário representado por valores numéricos, por exemplo, **12, 23.3, 345** *e* **9.5**, considerando que esses valores numéricos estão dentro dos intervalos a_1, b_4, c_2 e e_1, respectivamente.

Esse modo de apresentação (valores contínuos) da solução é relevante para domínios cujas variáveis são eminentemente numéricas e os usuários desejam obter o valor exato das variáveis que compõe o cenário ótimo, o que é o caso da aplicação principal desta tese, a qual é discutida no próximo capítulo.

Para realizar o processo de descoberta dos cenários com valores contínuos (numéricos), fazemos uso novamente de um algoritmo genético, chamado aqui de *AG_Multivariado*, cuja função de aptidão é obtida a partir dos dados. Essa função é calculada por meio de uma regressão de múltiplas variáveis realizada sobre os atributos da rede (Dillon e Goldstein, 1984); (Hair et al., 1998). Essa análise, no entanto, é realizada sobre os dados, mas considerando para a mesma apenas as instâncias da base de dados cujos atributos estejam compreendidos entre as faixas encontradas pelo AG explicado na seção 5.2.2.

O modelo de análise multivariada empregado é o de regressão múltipla, visto que estamos manipulando um conjunto de variáveis e desejamos saber a correlação existente entre as variáveis de evidências *E* (independente) e a variável meta *X* (dependente).

A técnica de regressão múltipla estabelece um modelo específico de análise multivariada, cuja principal finalidade é obtenção de uma relação entre uma variável específica (dependente) e as demais variáveis que descrevem o domínio (variáveis independentes), estabelecendo o grau de correlação que estas possuem para com a variável dependente. É por intermédio dessa função de correlação que é medida a aptidão dos indivíduos do *AG_Multivariado.*

Para tanto, podemos realizar o cálculo da correlação por meio da equação 5.3.

$$X_i = C_0 + C_1 E_{1i} + C_2 E_{2i} + \cdots + C_k E_{ki} + u_i \qquad \text{Equação 5.3}$$

na qual

X_i é a variável meta dependente;

E_{ki} são *as K* variáveis de evidência;

C_k são os coeficientes da regressão a serem estimados; e

u_i são os erros aleatórios.

O sistema geral de regressão múltipla pode ser então representado pelo conjunto de equações 5.4.

$$\begin{aligned} X_1 &= C_0 + C_1E_{11} + C_2E_{12} + \cdots + C_kE_{1k} + u_1 \\ X_2 &= C_0 + C_1E_{21} + C_2E_{22} + \cdots + C_kE_{2k} + u_2 \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ X_n &= C_0 + C_1E_{n1} + C_2E_{n2} + \cdots + C_kE_{nk} + u_n \end{aligned} \qquad \text{Equação 5.4}$$

Esse conjunto de equações pode ser reescrito como um sistema matricial, representado na equação 5.5.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & E_{11} & E_{12} \cdots E_{1k} \\ 1 & E_{21} & E_{22} \cdots E_{2k} \\ \vdots & \vdots & \vdots \quad \vdots \\ 1 & E_{n1} & E_{n2} \cdots E_{nk} \end{bmatrix} \times \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_k \end{bmatrix} + \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_n \end{bmatrix} \qquad \text{Equação 5.5}$$

Deduz-se da equação 5.5 que $X = EC + u$, em que:

X é um vetor coluna, com dimensão $n \times 1$ de observações ou registros da variável meta *X*;

E é uma matriz de ordem $n \times k$, ou seja, *n* observações e *k* variáveis, onde a primeira coluna representa o intercepto C_0;

C é um vetor coluna com $k \times 1$ parâmetros desconhecidos;

u é um vetor coluna de erros aleatórios com $n \times 1$.

Essa especificação objetiva gerar os valores dos parâmetros do vetor *C*. Para isto, emprega-se o método dos mínimos quadrados ordinários (MQO). Ele permite gerar estimativas para os parâmetros *C* com menor variância e não-viesados. O método de MQO

consiste em minimizar a soma dos erros quadrados da regressão estimada (equação 5.6), de tal forma que este seja o menor possível.

$$\min \sum_{i}^{n} u_i^2 = \sum_{i}^{n} \left(X_i - \hat{X}_i \right)^2 \qquad \text{Equação 5.6}$$

O resultado da aplicação do método de MQO ao modelo é dado pela equação 5.7.

$$C = \left(E^t E\right)^{-1} \times E^t X \qquad \text{Equação 5.7}$$

Com base nos valores de *C*, obtidos a partir da equação 5.7, é possível estabelecer a função de aptidão para os indivíduos do *AG_Multivariado*. Por exemplo, no caso da RB da figura 5.5 (a), considerando o valor meta X_1=145, o indivíduo (E_1=**12,** E_2=**23.3,** E_3=**345,** E_4= **9.5**), representado por uma *n-tupla* da base de dados, seria avaliado (aptidão) por meio da expressão:

$$X_1 = C_0 + 12C_1 + 23{,}3C_2 + 345C_3 + 9{,}5C_4 \qquad \text{Equação 5.8}$$

O algoritmo genético é então executado, com base da função de aptidão (equação 5.8), obtendo-se assim os valores, para cada um dos atributos, que compõem a configuração de valores de *E* que propiciam a obtenção de X_1 (no exemplo, 145). Destacando novamente que os indivíduos (instância da base de dados) avaliados pela função de aptidão são apenas aqueles que fazem parte das faixas de valores das variáveis que compõem o cenário encontrado pelo AG principal da estratégia de descoberta de cenários (figura 5.3).

Com o intuito de mostrar a lógica de funcionamento do *AG_Multivariado*, é mostrado, na figura 5.13, o módulo que permite a obtenção dos valores contínuos das variáveis que compõem o cenário descoberto pelo AG principal.

1. *AG_Multivariado* (*c, B*)
2. /* retorna o cenário, com valores contínuos, que melhor contribui para obtenção do valor meta para uma variável do domínio de aplicação */
3. // argumentos: *c* // cenário descoberto por DESCOBERTA_CENÁRIO (*rb*, *B*)
4. *B* // base de dados utilizada para a geração de *rb*
5. *população* ← POPULAÇÃO_INICIAL_ALEATÓRIA;
6. repita
7. *nova_população* ← *conjunto vazio*
8. para *i* ← 1 até TAMANHO(*população*) faça
9. **se (TUPLA(B) ⊂ *c*)**
10. *a* ← SELEÇÃO(*população*, **FUNÇÃO_APTIDÃO(REGRESSÃO_MÚLTIPLA(TUPLA(B))**
11. *b* ← SELEÇÃO(*população*, **FUNÇÃO_APTIDÃO(REGRESSÃO_MÚLTIPLA(TUPLA(B))**
12. se (TAXA_CROSSOVER é atendida)
13. *filho_ab* ← CROSSOVER(*a*,*b*)
14. se (TAXA_MUTAÇÃO é atendida)
15. *filho_ab* ← MUTAÇÃO(*filho_ab*)
16. incluir *filho_ab* em *nova_população*
17. fim se
18. fim para
19. *população* ← *nova_população;*
20. até algum indivíduo (cenário ótimo *c*) está adaptado o suficiente ou até ser obtido um número *n* de gerações
21. retornar o melhor cenário em população, de acordo com **REGRESSÃO_MÚLTIPLA**

Figura 5.13. Módulo de descoberta de cenários com valores contínuos.

Pode-se notar na figura 5.13, que o *AG_Multivariado* seleciona, para efeito da valoração de aptidão dos seus indivíduos, apenas as instâncias (TUPLAS) da base de dados *B* compreendidas (contidas) no cenário *c*, encontrado pelo AG principal. Para realizar a interação (chamada) entre o AG principal e o *AG_Multivariado*, é necessário modificar a última linha de DESCOBERTA_CENÁRIO (rb), conforma a seguir:

18. retornar o melhor cenário em população, de acordo com *AG_Multivariado*(*c*,B)

Para efeito de parametrização do *AG_Multivariado*, foram considerados os seguintes aspectos, relacionados na tabela 5.3.

Tabela 5.3. Parâmetros utilizados nos *AG_Multivariado*.

Parâmetro	Valores
População inicial	100 indivíduos
Número de gerações	1000 gerações
Seleção	método da roleta
Crossover	cruzamento de um ponto
Taxa de *Crossover*	98%
Taxa de mutação	0,1%

Além disso, foi empregada a representação binária, diferentemente do AG principal, que utilizou uma representação decimal dos cromossomos, conforme observado em 5.2.2.

5.2.5. Descoberta de Cenários – Mais de uma variável meta

Em alguns domínios de aplicação é interessante prover análises de cenários que estejam relacionados à obtenção de mais de um valor meta, isto é, que possuam mais de uma variável de consulta. Por exemplo, a estratégia poderia ser aplicada para a descoberta do cenário climático ideal para alcançar um valor meta de consumo de energia residencial e industrial.

Essa abordagem é implementada pelo método híbrido desenvolvido, atribuindo-se um peso para cada uma das variáveis de consulta.

Desse modo, a função de aptidão dos indivíduos (cenários) deve ser modificada, de acordo com a equação 5.9.

$$f = \sum_{i}^{n} w_i P(x_i \mid E) \qquad \text{Equação 5.9}$$

na qual

w_i é o peso associado a cada uma das metas (valores das variáveis metas x_i), normalizados $\sum_{x_i \in X} w_i = 1$; e

E é o conjunto de evidências (cenário).

Para entender a função ponderada de aptidão (equação 5.9), considere o exemplo da RB da figura 5.5(a) e a intenção de encontrar o cenário ótimo para obtenção dos valores meta $B=b_1$ e $D=d_2$. Assim, precisamos encontrar o valor de A, C e E que melhor reflitam às

metas (b_1 e d_2), considerando o peso, estabelecido *a priori* pelo especialista. Caso as metas possuam os mesmos pesos é atribuído para o peso de cada meta o valor *1/n*, no qual *n* está associado ao número de variáveis metas.

No exemplo, se considerarmos b_1 e d_2 com o mesmo peso (*1/2*), a função de avaliação da aptidão do indivíduo (a_2, c_1, e_3) pode ser representada da seguinte forma:

$$f = \frac{1}{2} P(b_1 \mid a_2, c_1, d_1, e_3) + \frac{1}{2} P(d_2 \mid a_2, b_1, c_1, e_3) \qquad \text{Equação 5.10}$$

O algoritmo genético é então executado, com base na função de aptidão (equação 5.9), obtendo-se assim os estados, para cada um dos atributos, que compõem a configuração de valores de *E* que propiciam a obtenção de X_1 (no exemplo, b_1 e d_2).

5.2.6. Condição de parada do algoritmo genético via critério subjetivo de avaliação do cenário descoberto

Como forma de considerar o conhecimento de fundo do especialista do domínio, especialmente relacionadas às expectativas do usuário quanto aos cenários descobertos, bem como estabelecer um critério de parada para o AG empregado na estratégia de descoberta de cenários, foram adotadas medidas subjetivas de avaliação desses cenários.

A idéia central é considerar, para efeito da descoberta de cenários, determinadas crenças que o usuário possui sobre o domínio e medir a influência que os padrões descobertos (cenários) exercem sobre essas crenças, de acordo com grau de inesperabilidade do padrão descoberto.

Para isso, foi empregada a abordagem bayesiana. Nessa abordagem, o grau de crença α é medido pela probabilidade condicional $P(\alpha|\xi)$, na qual ξ representa alguma evidência do usuário sobre o domínio associada à crença α. Dada uma nova evidência *E*, representada na estratégia desenvolvida aqui por um cenário, é necessário atualizar o grau da crença para $P(\alpha| \xi, E)$, calculado a partir do emprego do teorema de Bayes (equação 3.15), conforme a equação 5.11.

$$P(\alpha \mid \xi, E) = \frac{P(E \mid \alpha, \xi) P(\alpha \mid \xi)}{P(E \mid \alpha, \xi) P(\alpha \mid \xi) + P(E \mid \neg\alpha, \xi) P(\neg\alpha \mid \xi)} \qquad \text{equação 5.11}$$

Assim, considerando como exemplo o domínio representado pela RB da figura 5.5(a) e a crença do usuário α_1 = "quanto maior o valor de *B* e *C*, maior a possibilidade da meta d_1 ser atingida", com um grau de crença (medida via método de inferência da RB), por exemplo, igual a 0,78 ($P(\alpha_1|\xi)=P(d_1|b_4,c_2)$, onde b_4 é o maior valor de *B* e c_2 é o maior de *C*). Além disso, considere uma nova evidência (cenário) E_1, representado por (a_2, b_3, c_2, e_1). É possível associar as funções com a crença α_1, calculando-se as probabilidades condicionais $P(E_1|\alpha_1,\xi)$ e $P(E_1|\neg\alpha_1,\xi)$. Assumindo que essas probabilidades são 0,57 e 0,48, respectivamente, e utilizando esses valores na equação 5.11, teríamos $P(\alpha_1|\xi,E_1)=0,8$.

O grau de interesse de um cenário (padrão descoberto) é função de como o mesmo influencia na crença do usuário. Isto é, quanto maior for o impacto desse padrão sobre a crença do usuário, maior deve ser o grau de interesse desse padrão (Silberschatz e Tuzhilin, 1996). Para quantificar esse impacto, podem-se estabelecer pesos para cada crença e calcular o grau de interesse de um determinado cenário a partir da equação 5.12.

$$I(c,B,\xi)=\sum_{\alpha_i\in B} w_i \,|\, d(\alpha_i \,|\, c,\xi)-d(\alpha_i \,|\, \xi)\,| \qquad \text{equação 5.12}$$

no qual,

I é o grau de interesse do cenário *c*;

B é o conjunto de crenças α_i;

W_i é o peso associado a cada evidência α_i normalizados $\sum_{\alpha_i\in B} w_i = 1$; e

ξ é a evidência prévia utilizada pelo usuário como base de sua crença inicial

A partir da equação 5.12, é possível avaliar o quanto o grau de crença varia de acordo com o impacto do cenário *c* sobre essa crença. Por exemplo, se considerarmos que o peso de cada crença α_i do conjunto de crenças *B* tem o mesmo valor (isto é *1/n*, no qual *n* é o número de crenças de *B*), o exemplo da crença α_1 = "quanto maior o valor de B e C, maior a possibilidade da meta d_1 ser atingida", com $P(\alpha_1|\xi)=0,78$, e que essa evidência foi modificada, com a descoberta de um novo cenário E_1, para $P(\alpha_1|\ \xi,\ E_1)=0,8$, é possível obter, substituindo esses valores na equação 5.12, o grau de interesse do padrão E_1, para a crença α_1, sendo o mesmo $I(E_1,\alpha_1,\xi)$=0,02.

Para estabelecer o critério de parada do AG, é possível definir um valor mínimo para o grau de interesse do padrão. Por exemplo, o AG poderia ser repetido (número de gerações) até que o grau de interesse do padrão descoberto fosse superior a 0,2. Sendo, desse modo, necessárias as seguintes modificações no algoritmo da figura 5.3, a seguir reescrito, na figura 5.14, com tais alterações em destaque.

```
1.  DESCOBERTA_CENÁRIO (rb, crença)
2.  /* retorna o cenário que melhor contribui para obtenção de valor meta para uma
    variável do domínio de aplicação */
3.  // argumentos: rb //rede bayesiana que codifica a distribuição conjunta P(X1, X2,
    ...Xn)
                    crença: crença sobre um determinado aspecto do domínio,
                                considerando o valor meta que se deseja obter
4.  população ← POPULAÇÃO_INICIAL_ALEATÓRIA;
5.  repita
6.    nova_população ← conjunto vazio
7.    para i ← 1 até TAMANHO(população) faça
8.      a ← SELEÇÃO(população, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(rb))
9.      b ← SELEÇÃO(população, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(rb))
10.     se TAXA_CROSSOVER é atendida
11.       filho_ab ← CROSSOVER(a,b)
12.     se TAXA_MUTAÇÃO é atendida
13.       filho_ab ← MUTAÇÃO(filho_ab)
14.     GRAU_INTERESSE(c, filho_ab)
15.     incluir filho_ab em nova_população
16.   fim para
17.   população ← nova_população;
18.  até algum indivíduo (cenário) possuir um grau de interesse satisfatório
19. retornar o melhor cenário em população, de acordo com FUNÇÃO_APTIDÃO
```

Figura 5.14. Algoritmo do processo de descoberta de cenários, considerando como critério de parada o grau de interesse do cenário descoberto.

Decorre-se da figura 5.14, que para que se possa estabelecer um critério de parada para o AG, é necessário definir, de modo antecipado, a crença *c* do usuário para a obtenção de uma meta, com base nas demais variáveis do domínio, bem como um valor aceitável para o grau de interesse, medido de acordo com a inesperabilidade dos cenários descobertos e calculado por GRAU_INTERESSE, mediante o emprego da equação 5.12.

A principal motivação para implementação desse critério para o AG da estratégia de descoberta de cenários é a possibilidade de embutir no processo de busca dos cenários o conhecimento a priori do especialista sobre o domínio de aplicação.

5.3. CONSIDERAÇÕES FINAIS

Neste capítulo, foi descrito o contexto em que está enquadrada esta tese, dentro de um conjunto de estratégias elaboradas com o objetivo precípuo de estender as potencialidades das RBs, no que concerne aos processos de aprendizado e inferência dessas redes. Em adição, foi possível compreender a estratégia para obtenção de cenários, a partir da elaboração de um método híbrido, utilizando RBs e AG, bem como de todas as variações desse método. Uma das principais características do método em questão é prover um novo mecanismo de inferência em RBs, estendendo com isso seu poder de interpretabilidade.

Destaca-se que, apesar das abordagens serem apresentadas de maneira isolada, as mesmas podem ser combinadas de acordo com as especificidades do domínio de aplicação. Desse modo, por exemplo, é possível obter as variáveis que exercem maior influência sobre duas variáveis meta ou o cenário com valores contínuos que melhor explique a obtenção dos valores meta de duas ou mais variáveis, considerando como critério de parada do AG a medida do grau de interesse do cenário descoberto.

Em razão de uma das principais motivações para elaboração desta estratégia advir das necessidades demandadas do setor elétrico, particularmente dos projetos de pesquisa, desenvolvidos no âmbito do Laboratório de Planejamento de Redes de Alto Desempenho (LPRAD), é demonstrada, no próximo capítulo, a aplicação do método desenvolvido, com vistas, principalmente, à descoberta de cenários favoráveis à obtenção de um valor meta de consumo de energia elétrica.

6. ESTUDO DE CASO: DESCOBERTA DE CENÁRIOS SÓCIO-ECONÔMICOS E CLIMÁTICOS PARA OTIMIZAÇÃO DO CONSUMO DE ENERGIA ELÉTRICA

6.1. CONSIDERAÇÕES INICIAIS

Conforme evidenciado no capítulo 4, são latentes as demandas do setor elétrico para as soluções dos problemas emergentes nessa área. Muitas dessas soluções estão voltadas para a resolução de uma classe de problemas típicos de otimização, bem como outras são próprias para serem equacionadas por métodos que envolvam a modelagem de dependência de suas variáveis.

Neste capítulo, são abordadas as principais motivações e a contextualização da estratégia desenvolvida, bem como a aplicação do método para o apoio ao processo de tomada de decisão do setor elétrico, fundamentalmente relacionada ao mercado de energia elétrica.

6.2. MOTIVAÇÃO E CONTEXTUALIZAÇÃO DA ESTRATÉGIA DESENVOLVIDA

As investigações aqui descritas originaram-se a partir das demandas do projeto de pesquisa "Predict - Ferramenta de Suporte à Decisão para Predição de Cargas de Sistemas Elétricos", financiado pela "Agência Nacional de Energia Elétrica do Brasil (ANEEL)" e que está em curso desde setembro de 2004.

Esse projeto, realizado em parceria com o Governo do Estado do Pará e a Concessionária de Energia Elétrica do Estado do Pará (CELPA), visa, basicamente, projetar e implementar um sistema de suporte à decisão, utilizando métodos matemáticos e de inteligência computacional, para prever as necessidades de compra de energia no mercado futuro, bem como para realizar inferências sobre a situação do sistema elétrico, a partir de dados históricos de consumo e suas correlações com dados sócio-econômicos e climáticos.

O Predict pauta-se na certeza de que a previsão de carga (carga elétrica em MW) é uma estratégia primordial dos sistemas elétricos. É baseada nessa previsão, que se planejam e operam esses sistemas de forma confiável e segura. Tipicamente, em previsão de carga, pretende-se definir qual o consumo de energia futuro de uma dada região de modo, por exemplo, a projetar ou adequar o sistema elétrico para atender esses consumidores quando essas demandas se concretizarem no futuro. Assim, baseado no histórico das medidas do sistema, deseja-se obter uma prospecção das necessidades futuras.

Nesse horizonte, via de regra, devem ser realizados estudos para medir o impacto que diversas outras variáveis aleatórias (temperatura, umidade, fatores sócio-econômicos, dentre outros) exercem sobre o consumo, de tal sorte que seja possível prever cenários cuja operação do sistema elétrico seja econômica, segura, confiável e de boa qualidade.

Dessa forma, a previsão de consumo e a correlação de algumas variáveis exógenas ao sistema elétrico, especificamente associadas a fatores climáticos e sócio-econômicos, serviu de base para a criação do Predict, o qual, em sua primeira fase, realizada no período de 2004 a 2006, utilizou métodos de regressão e redes neurais artificiais, para efetuar as previsões e RBs para modelar as correlações supracitadas.

Entretanto, ao longo do desenvolvimento do Predict, foram levantadas, pelos especialistas (gerentes e engenheiros), uma série de novas inferências necessárias ao planejamento e a operação dos sistemas elétricos. Dentre as novas demandas para a melhoria do Sistema de Suporte à Decisão, podem ser destacadas:

- Aplicação das ferramentas desenvolvidas no Predict para predição e modelagem das correlações entre os fatores sócio-econômicos e climáticos e o consumo de energia nos outros estados atendidos por outras concessionárias do Grupo Rede Energia (CELTINS - Tocantins, CEMAT - Mato Grosso, ENERSUL - Mato Grosso do Sul, REDESUL - algumas regiões do Norte do Paraná e Sul de São Paulo), do qual também faz parte a Concessionária CELPA (Pará);
- Geração de indicadores que influenciam o desempenho futuro do Sistema Elétrico, como os mecanismos que identifiquem a correlação, ao longo do tempo, entre as varáveis sócio-econômicas e climáticas, com este consumo;

- Consideração nas análises propostas do potencial de economia de energia (energia conservada) com vistas ao planejamento de mercado;
- Descoberta de cenários sócio-econômicos e climáticos propícios à obtenção de de um valor-meta de consumo;

Essas demandas evidenciaram a importância da implementação de novos métodos, a fim de dar respostas ainda mais abrangentes à busca de padrões de consumo de energia elétrica. Para isto, um conjunto de novas funcionalidades foi proposto para ser acrescentado ao Predict, o que está sendo realizado na continuidade do projeto, cuja renovação foi aprovada pela ANEEL e está em curso desde 2007.

A partir dessas novas necessidades, é possível diagnosticar que o atendimento das mesmas passa pela compreensão e da mensuração da influência exercida por diversos fatores, elencados pelos especialistas do domínio, sobre o consumo de energia elétrica. Para isso foi eleito, como modelo básico de representação dessas correlações, as RBs. Entretanto, o formalismo básico destes modelos não foi suficientemente adequado para atender a todas essas necessidades, razão pela qual a estratégia de descoberta de cenários foi criada como resposta a essas demandas do setor elétrico.

Na seção 6.4, como forma de mostrar as contribuições da estratégia desenvolvida para o setor elétrico, são apresentadas as aplicações de suas diversas abordagens, principalmente voltadas para o apoio às decisões de compra de energia, fundamentada nos cenários de consumo estabelecidos pelo objeto de estudo central desta tese. À luz do indicativo da importância da estratégia para os procedimentos de comercialização de energia, por parte da concessionária de energia elétrica, e com intuito de esclarecer alguns dos pontos que fazem com que a precisão no acerto das previsões de consumo seja imperativa para o sucesso de suas transações comerciais, é apresentado, na seção seguinte, um breve cenário da comercialização de energia elétrica no País.

6.3. PROCESSO DE COMERCIALIZAÇÃO DE ENERGIA ELÉTRICA NO BRASIL

O Processo de Comercialização de Energia Elétrica no Brasil segue os parâmetros estabelecidos pela Lei nº 10848/2004, pelos Decretos nº 5163/2004 e nº 5.177/2004, que

instituiu a Câmara de Comercialização de Energia Elétrica (CCEE) e pela Resolução Normativa ANEEL nº 109/2004.

As relações comerciais entre os agentes (de geração, de distribuição e de comercialização de energia) participantes da CCEE são regidas predominantemente por contratos de compra e venda de energia, e todos os contratos celebrados entre os agentes no âmbito do Sistema Interligado Nacional devem ser registrados na CCEE.

A CCEE contabiliza as diferenças entre o que foi produzido ou consumido e o que foi contratado. As diferenças positivas ou negativas são liquidadas no Mercado de Curto Prazo e valorado ao PLD (Preço de Liquidação das Diferenças). Dessa forma, pode-se dizer que o mercado de curto prazo é o mercado das diferenças entre montantes contratados e montantes medidos.

Se por um lado, com a centralização das aquisições de energia por parte da CCEE, os riscos das distribuidoras aparentemente diminuíram, decorrente do fato das mesmas deixaram de ser responsáveis por viabilizar projetos de geração e, em consequência, por fazer as aquisições em volume suficiente para atender à demanda de suas áreas de atuação. Por outro, o novo modelo impõe um risco considerável nas estimativas de energia requerida para atendimento de suas áreas de concessão, posto que as mesmas são punidas caso cometam erros de previsão (Barros et al., 2009).

Soma-se a isso o fato de que o valor das punições é calculado com base em uma componente estocástica com alto grau de volatilidade - o PLD, sob a qual as distribuidoras não têm ingerência, visto que o PLD depende de vários fatores, tais como o dos níveis dos reservatórios e a expansão do sistema hidrotérmico. Face à variação do PLD, há um considerável risco na possibilidade de punição decorrente de erros na estimativa da demanda.

Previsões incorretas na compra de energia elétrica podem afetar negativamente as operações das concessionárias. As previsões abaixo do necessário, além de sujeitar essas empresas a multas, ainda as obrigam a celebrar contratos de energia de curto prazo, normalmente mais custosos, para atender a demanda dos seus consumidores. Caso sejam realizadas previsões acima do necessário, também decorrem penalidades da ANEEL. Além

do que nem sempre é possível repassar integralmente às tarifas, os custos advindos dessas previsões incorretas.

As regulamentações da ANEEL estabelecem que as distribuidoras de energia devam contratar antecipadamente, por meio de leilões públicos suas necessidades de energia. Essas normas também delineiam as condições gerais para o repasse dos volumes e preços de comercialização de energia. Caso a energia contratada, incluindo aquela comprada em leilões públicos, após aplicação do Mecanismo de Compensação de Sobras e Déficits de Energia (MCSD)[3] for inferior a 100% da necessidade de energia total, as empresas estarão sujeitas a multas. Se a energia contratada for superior a 100% e inferior a 103% da necessidade de energia total, pode-se repassar o volume total de energia comprada para os consumidores. Já no caso da energia contratada ser superior a 103% da necessidade de energia total, tem-se que assumir o risco entre a diferença de preço de compra nos leilões públicos e venda no mercado de curto prazo, não sendo permitido, ainda, repassar esses custos aos consumidores.

As incertezas as quais o mercado de energia elétrica está sujeito, mostrados nesta seção, elevam ainda mais a importância do uso de métodos inteligentes que possam identificar cenários de consumo de energia elétrica, tomando por base as informações endógenas e exógenas aos sistemas elétricos. Esses cenários podem ser úteis para a realização de contratos de compra de energia mais vantajosos no futuro, além de permitir que as entidades governamentais possam estabelecer políticas e investimentos para o desenvolvimento de certas regiões do Estado, face às correlações obtidas entre os dados de consumo e sócio-econômicos.

6.4. DESCOBERTA DE CENÁRIOS DE CONSUMO DE ENERGIA

Para ilustrar a aplicação da estratégia e todas as suas abordagens, nesta seção, serão mostradas cinco aplicações, a saber:

[3]O MCSD estabelece a cessão de montantes contratuais de energia entre distribuidoras com sobras de energia possam transferi-las para os distribuidores com déficits, mediante assinatura de termos de cessão. Ocorrerá anualmente o processamento do MCSD 4% decorrentes de outras variações de mercado, hipótese na qual poderá haver redução de ate 4% do montante inicial contratado nos leilões de energia proveniente de empreendimentos existentes.

- Descoberta do cenário sócio-econômico ótimo que explique a maximização do consumo, utilizando métodos exatos e aproximados de inferência;
- Identificação das variáveis que mais influenciam na obtenção de um valor meta de consumo.
- Descoberta do cenário sócio-econômico ótimo que corrobore com a obtenção de um valor meta de consumo total de energia, proposto pelo usuário. Adicionalmente, o cenário encontrado é apresentado com seus valores numéricos (contínuos);
- Descoberta do cenário sócio-econômico ótimo que explique a maximização do consumo residencial e comercial;
- Descoberta do cenário climático ótimo que propicie um valor meta do consumo total de energia, estabelecido pelo usuário, conforme o grau de interesse definido pelo mesmo;

Para os estudos de casos realizados nesta tese, são consideradas as bases de dados de consumo de energia elétrica de duas empresas do Grupo Rede Energia, são elas: CELPA (Pará) e ENERSUL (Mato Grosso do Sul). Os dados sócio-econômicos foram obtidos dos Governos Estaduais do Pará e do Mato Grosso do Sul e as informações climáticas do Instituto Nacional de Pesquisas Espaciais (INPE). A periodicidade dos dados é mensal e em razão de termos dados históricos do Governo do Pará somente a partir do ano 1999, foram consideradas, para efeito das análises do impacto dos dados sócio-econômicos, o consumo de energia a partir desse período, no que pese a CELPA possuir em seus registros dados consistentes desde 1989.

No que concerne aos dados da ENERSUL, o Governo do Mato Grosso do Sul disponibilizou dados consistentes desde 2000, o que fez com que considerarmos apenas os dados de consumo de energia elétrica a partir do ano de 2000, mesmo a ENERSUL tendo em suas bases registros das medidas de consumo, mês a mês, a partir de 1991.

Os impactos climáticos sobre o consumo de energia elétrica consideram os dados obtidos a partir do ano 2000, para ambas as companhias de energia elétrica.

Ressalta-se que em todas as aplicações do estudo do estudo de caso, os atributos das bases de dados de energia, climáticos e sócio-econômicos foram discretizados, por frequência, em dez faixas (estados).

Pelo fato da estratégia desenvolvida ter sido incorporada ao Sistema de Software Predict e estar sendo utilizada pelas Empresas do Grupo Rede Energia, é necessário compreender como é constituída a sua arquitetura, o que é feito na próxima seção.

6.4.1. Arquitetura Básica do Predict

Como forma de compreender melhor quais as fontes de dados utilizadas, bem como está estruturado o Sistema de Suporte à Decisão Predict, é apresentada na figura 6.1 a arquitetura simplificada do Sistema.

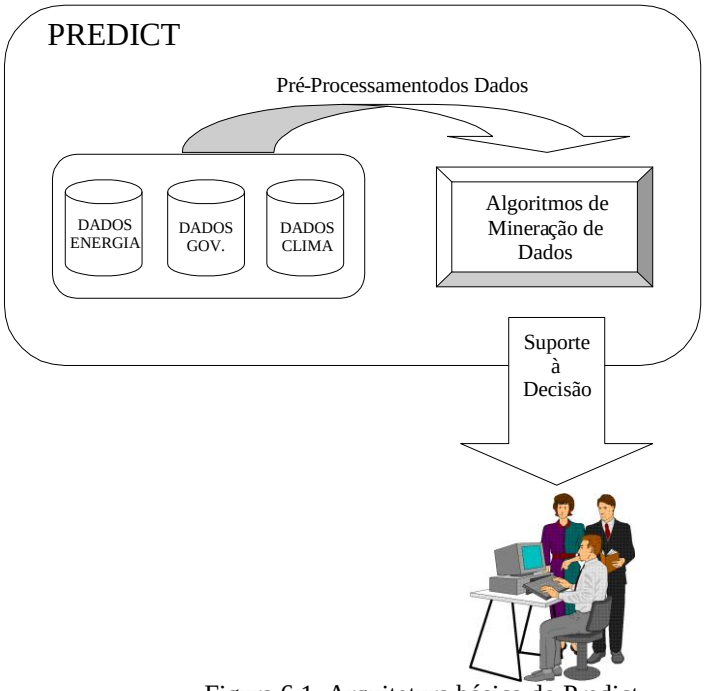

Figura 6.1. Arquitetura básica do Predict.

Os elementos desta arquitetura são divididos em módulos (subsistemas) separados, descritos as seguir:

- Gerência de Dados: contém os sistemas de aquisição e pré-processamento de dados utilizados no processo de Mineração de Dados.

- Gerência de Conhecimento: contempla o uso dos algoritmos de Mineração de Dados para descoberta de padrões.
- Gerência de Interface: engloba o uso racional de ferramentas de visualização das informações e/ou conhecimentos extraídos a partir dos dados, como forma de facilitar o entendimento, por parte dos usuários de níveis decisórios, dos resultados obtidos.

Em se tratando do módulo de Gerência de Dados, é preciso considerar as três fontes de dados, identificados na figura 6.1:

- Energia: dados oriundos da base de dados corporativa das concessionárias de Energia Elétrica do Grupo Rede Energia. No caso específico da CELPA e ENERSUL, utilizadas neste estudo de caso, foram consideradas as informações históricas do Sistema Comercial, constituída por 11 atributos (mês/ano, energia consumida residencial, comercial, industrial, rural, poder público, iluminação pública, serviço público, consumo próprio, consumo total faturado, energia requerida), medidos mensalmente. Para os estudos aqui apresentados, foram utilizados os dados de energia de janeiro de 1991 a agosto de 2009, no caso da ENERSUL e de janeiro de 1989 a agosto de 2009, no caso da CELPA.
- Governo: os dados de governo variam de acordo com os dados disponibilizados pelos Estados[4]. No caso do Pará, foram utilizados cinco atributos (valor total das receitas, valor médio do dólar, número de contratações nas indústrias de transformação e número de contratações na agroindústria), medidos mensalmente. No caso do Mato Grosso do sul, foram considerados 12 atributos (exportação de cereais, de sementes e frutos oleaginosos, de óleos vegetais, de açúcares, de minérios, de couros e peles, de madeira, de ferro e aço, arrecadação de ICMS, transferências da União, PIB (R$) e PIB (US$). Para os estudos aqui apresentados, foram utilizados os dados sócio-econômicos de janeiro de 2000 a agosto de 2009, no caso da ENERSUL e de janeiro de 1999 a agosto de 2009, no caso da CELPA.

[4] Os dados foram fornecidos pelas secretarias de planejamento dos estados do Pará e Mato Grosso do Sul.

- Climáticos: os dados climáticos consideram informações relativas à temperatura mínima, máxima, índice pluviométrico e umidade relativa do ar, perfazendo um total de quatro atributos. Foram consideradas para efetuar as correlações as médias mensais estaduais calculadas a partir dos pontos de coleta espalhados nos Estados do Pará e do Mato Grosso. Desse modo, as análises, realizadas como parte dos experimentos deste trabalho, foram executadas no âmbito estadual, no que pese algumas RBs geradas pelo Predict também efetuarem análises da correlação dos fatores climáticos com o consumo especificamente nos municípios em que há ponto de coleta dos valores climáticos. Para os estudos aqui apresentados, foram utilizados os dados climáticos de janeiro de 2000 a agosto de 2009, de ambos os estados.

Destaca-se que a escolha dos atributos climáticos e sócio-econômica se deu por meio das intervenções dos especialistas do domínio (analistas de mercado e engenheiros das concessionárias).

No módulo de Conhecimento, são destacadas duas linhas de análise, elencadas nos itens abaixo:

- Previsão: responsável pela previsão do consumo de energia faturado (e suas várias classes) e da energia requerida de médio (1 a 2 anos) e longo prazos (mais de 2 anos). Para realizar essa tarefa, são utilizados métodos de regressão matemáticos, redes neurais artificiais, além de métodos híbridos neuro-genéticos.

- Correlação: responsável pela codificação da correlação entre os dados sócio-econômicos, climáticos e de energia elétrica. A influência climática e sócio-econômica sobre o consumo de energia elétrica é realizada por meio das redes bayesianas e por métodos híbridos que combinam RB com cadeias de Markov e Algoritmos Genéticos. A estratégia, objeto das investigações realizadas nesta Tese, emprega as análises demandadas por esse módulo como estudo de caso.

Por fim, o Módulo de Interface, que apresenta os resultados de modo amigável para os usuários do Sistema Predict. Na figura 6.2, é apresentada a interface do módulo de geração de cenário. O exemplo em questão diz respeito ao cenário ideal para a

maximização do consumo comercial, utilizando os dados climáticos e do consumo faturado de energia, por classe, da concessionária ENERSUL.

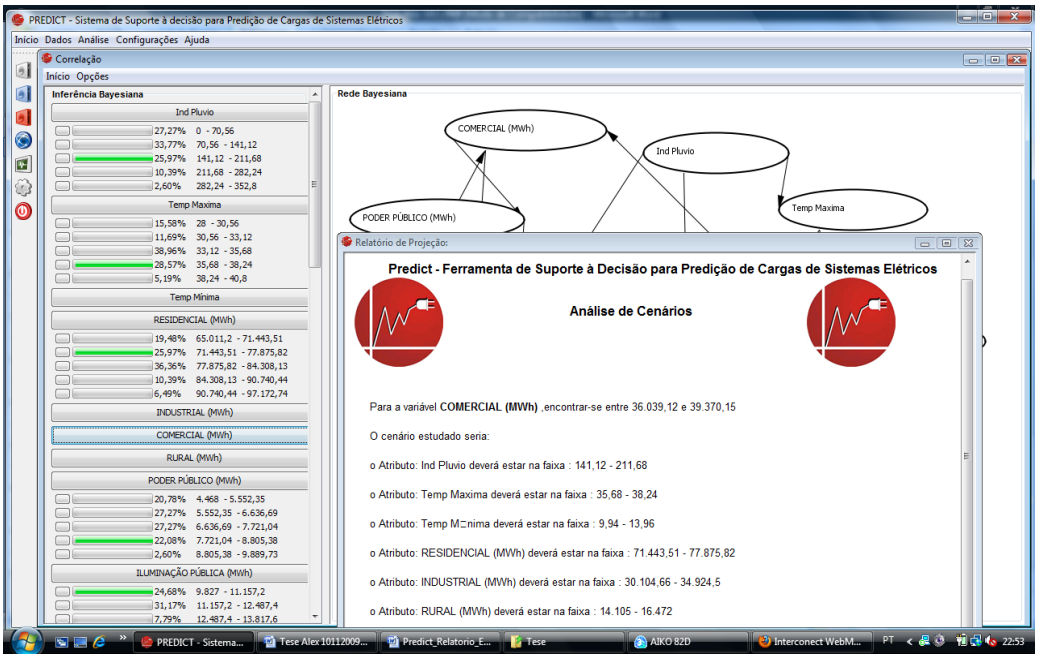

Figura 6.2. Exemplo da interface gráfica do Predict.

Compreendida a arquitetura do Predict, são apresentados, nas seções 6.4.2 a 6.4.5, os resultados do emprego da estratégia de descoberta de cenários e suas variações no contexto das bases de dados do Sistema Predict. Em todas as situações propostas nessas seções, os cenários são obtidos sobre as RBs geradas a partir dos dados do Predict, utilizando o algoritmo de busca e pontuação K2 (Cooper e Herskovitz, 1992).

6.4.2. Descoberta de cenários utilizando métodos exatos e aproximados de inferência bayesiana

Conforme abordado na seção 5.2.2, a estratégia desenvolvida trabalha tanto com métodos exatos quanto com métodos aproximados para quantificar a aptidão dos indivíduos do AG. Cabe ressaltar que por mais que haja pequenas variações nos valores das probabilidades marginais dos nós das redes, ocorridas com o uso de métodos aproximados, quando comparados aos métodos exatos, os cenários gerados foram similares nos experimentos realizados.

Neste contexto, a idéia central da aplicação proposta nesta seção, é encontrar o cenário sócio-econômico que maximiza o consumo de energia residencial da concessionária ENERSUL. A RB gerada para essa análise, considera os 18 atributos (sócio-econômico e de consumo de energia por classe) e a topologia apresentada na figura 6.3.

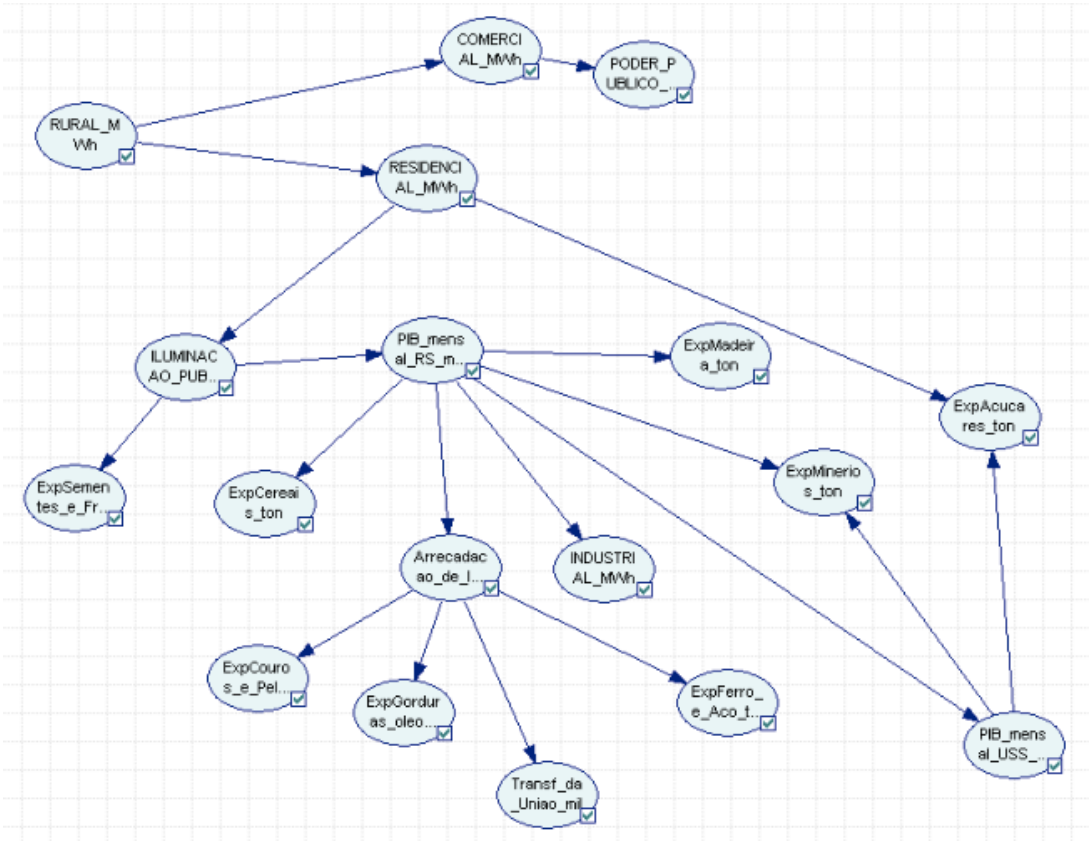

Figura 6.3. Rede bayesiana gerada para medir as correlações entre os dados sócio-econômicos e o consumo de energia elétrica por classe.

Para isso, são empregados, como exemplo, o método exato Árvore de Junção e o método aproximado *AIS Sampling*. Lembrando que poderiam ser utilizados outros métodos de inferências, quer sejam exatos ou aproximados.

Utilizando-se o método de Árvore de Junção foi encontrado o seguinte cenário (7, M, 7, 5, 3, 5, 8, 1, 2, 1, 9, 9, 8, 2, 4, 7, 3, 10, para a obtenção de M, com máxima probabilidade, isto é, RESIDENCIAL_MWh = 10). A probabilidade máxima obtida foi de 0,833617. Já utilizando o método aproximado *AIS Sampling*, foi possível obter a probabilidade máxima de 0,822175 de ocorrência da meta (consumo máximo de energia residencial).

Analisando os resultados obtidos com os métodos de inferência exato e aproximado, podemos constatar que, no que pese as diferenças existentes nas configurações dos estados das variáveis, as probabilidades máximas de obtenção do valor meta estão muito próximos. Isso se deve a baixa complexidade da RB Enersul. Desse modo, como não há impedimento no emprego de métodos exatos nas RBs geradas para o estudo de caso abordado nesse Capítulo, bem como evitarmos a variabilidade dos resultados dos métodos de inferência aproximados, iremos adotar, em todos os demais experimentos, o método exato, mas especificamente o de Árvore de Junção.

6.4.3. Identificação das variáveis de maior influência sobre a maximização do consumo residencial

Para encontrar as variáveis que mais influenciam na maximização da probabilidade de se obter o máximo consumo residencial da ENERSUL, utilizamos a variação do AG que efetua a busca de cenários discutida na seção 5.2.3. Portanto, aplicando tal mecanismo, é possível obter a seguinte configuração dos indivíduos do AG, os quais são mostrados na tabela 6.1 (por simplificação, são apresentados apenas os 10 indivíduos com maior aptidão).

Tabela 6.1. Indivíduos gerados após 1000 gerações do AG, utilizado para encontrar os atributos que mais influenciam na obtenção da meta (consumo residencial máximo, i.e. RESIDENCIAL_MWh=10).

Indivíduo																		**Aptidão**
A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	***P(X|E,Y)***
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	1	0	9	0	0	0	0	0	0	5	0	0	0	0,780233
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0,784660
7	M	7	0	0	1	0	9	0	0	0	9	0	0	5	0	0	0	0,780233
7	M	7	0	0	0	0	10	0	7	3	0	0	0	0	0	0	4	0,775356
7	M	7	0	0	0	0	10	0	7	3	0	0	0	0	0	0	4	0,775356

Analisando a aplicação da estratégia desenvolvida, considerando o conjunto total de variáveis de evidência, o qual encontrou a mais provável configuração (7, M, 7, 5, 3, 5, 8, 1, 2, 1, 9, 9, 8, 2, 4, 7, 3, 10) para obtenção de M (RESIDENCIAL_MWh = 10), mostrada na seção anterior, em relação à observação das variáveis mais influentes, pode-se notar (em destaque na tabela 6.1) que as variáveis *RURAL_MWh (na tabela 6.1, variável A1)*, *ILUMINACAO_PUBLICA_MWh (A3)* e *ExpCouros_e_Peles_ton (A12)*, quando

instanciadas com os valores *RURAL_MWh = 7*, *ILUMINACAO_PUBLICA_MWh* = 7, e *ExpCouros_e_Peles_ton* = 9, tem maior probabilidade de obtenção do valor meta (RESIDENCIAL_MWh = 10) – *0,784660*. Pode-se dizer, portanto, que as variáveis *RURAL_MWh*, *ILUMINACAO_PUBLICA_MWh* e *ExpCouros_e_Peles_ton*, quando combinadas, exercem maior impacto na meta a ser alcançada.

6.4.4. Descoberta do cenário sócio-econômico (valores contínuos) ótimo que corrobore com a obtenção de um valor meta de consumo

Para ilustrar as potencialidades da estratégia de descoberta de cenários com valores contínuos das variáveis de evidência, considere o processo de descoberta do cenário sócio-econômico ótimo para a obtenção de um valor máximo do consumo total de energia faturado pela Concessionária de Energia Elétrica do Estado do Pará (CELPA).

Para isso, são utilizados os atributos sócio-econômicos considerados para a análise da correlação com o consumo total faturado, são eles: número de contratações no setor das indústrias de transformação (*cont_ind*), contratações no setor de agropecuária (*cont_agro*), o valor do total da receita (*val_rec*) e o valor do dólar (*val_dol*). Utilizando o K2, foi obtida a RB, ilustrada na figura 6.4.

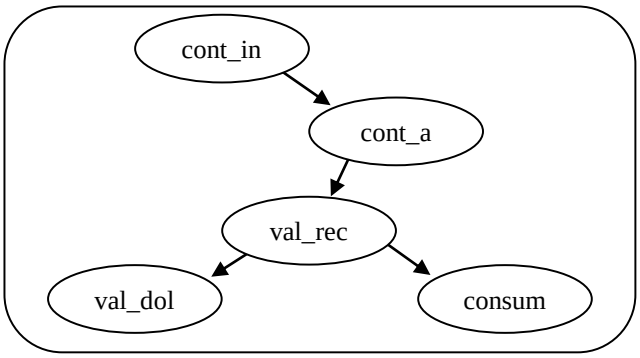

Figura 6.4. Rede Bayesiana gerada a partir do K2.

Na RB montada, todos os atributos foram discretizados em dez estados (faixas), segundo a freqüência de seus valores, permitindo-nos verificar a probabilidade associada a cada uma delas, assim como as probabilidades condicionais existentes entre as variáveis. Vale sublinhar que o número de estados das variáveis poderia ser diferente entre si, bem como em valores maiores ou menores que 10 faixas.

Uma vez montada a RB, a próxima etapa consiste em, se valendo dos dados disponíveis na RB, buscar nos atributos da rede as faixas que maximizariam o consumo de energia. Nesta etapa, fazemos uso do algoritmo genético apresentado na seção 5.2.2. Para efeito de aplicação, foi utilizado o método exato Árvore de Junção como função de aptidão dos indivíduos da população.

Assim, utilizando-se como critério de parada o número de gerações, no caso mil gerações, foi encontrado o cenário ótimo que maximiza o consumo, mostrado na figura 6.5.

2	1	7	4

Figura 6.5. Cenário ótimo para maximização do consumo total faturado.

Isto é, as variáveis *cont_ind* na faixa 2, *cont_agro* na faixa 1, *val_rec* na 7 e *val_dol* na 4. Recordando que cada indivíduo, no formato apresentado na figura 6.5, é, para efeito de sua classificação, submetido ao módulo de inferência Bayesiana para verificar a probabilidade em que o atributo consumo seria maximizado, obtendo assim, ao término das iterações, a melhor configuração possível de inferências na RB, para a maximização do consumo.

No entanto, teríamos ao término deste passo (após a análise pelo algoritmo genético) apenas as respectivas faixas de valores para essa maximização, ao invés do valor contínuo para cada atributo, que é o que buscamos.

Para tal, fazemos uso novamente de um algoritmo genético *AG_Multivariado*, explicado na seção 5.2.3. Conforme também sinalizado em 5.2.3, o *AG_Multivariado* tem seu espaço de busca reduzido às instâncias da base de dados que se localizam dentre as faixas encontradas na etapa anterior. Ou seja, *cont_ind* na faixa 2, *cont_agro* na faixa 1, *val_rec* na 7 e *val_dol* na 4. Dessa forma, obtemos uma equação com uma boa representatividade (R^2 = 0,9039) sobre o domínio.

Como também relatado na seção 5.2.3, o cálculo da função de aptidão dos indivíduos do *AG_Multivariado* é estimado por meio de uma regressão de múltiplas variáveis, que estima os coeficientes que possam estabelecer o maior grau de correlação entre a variável meta *Y* (consumo total de energia faturado) e as variáveis sócio-econômicas. Assim, após a execução do *AG_Multivariado*, obtemos a equação 6.1.

 Equação 6.1

na qual:

Y representa o consumo de energia e;

X_1, X_2, X_3 e X_4 representam os valores dos atributos *cont_ind*, *cont_agro*, *val_rec* e *val_dol*, respectivamente.

Com base na equação 6.1, o algoritmo genético é então utilizado, obtendo-se assim os valores, para cada um dos atributos, que maximizam o consumo total de energia faturado.

Assim, após as iterações do *AG_Multivariado*, no caso mil gerações, obtemos o seguinte cenário, mostrado na tabela 6.2, propício à ocorrência do consumo total de energia faturado máximo, equivalente a 405.760MWh.

Tabela 6.2. Valores dos atributos para maximização do valor do consumo.

Atributo	**Valor**
cont_ind	5.380
cont_agro	3.357
val_rec	R$ 100.752.576,00
val_dol	R$ 2,861

Tanto o AG que utiliza a RB da figura 6.4 como função de custo, quanto o *AG_Multivariado*, foram basicamente parametrizados de acordo com o estabelecido na tabela 6.3. Foram realizados experimentos especificando outros valores para os parâmetros da tabela 6.3, entretanto os resultados não sofreram alterações que merecessem registro.

Tabela 6.3. Parâmetros utilizados nos AGs.

Parâmetro	**Valores**
População inicial	50 indivíduos
Número de gerações	1000 gerações
Seleção	método da roleta
Crossover	cruzamento de um ponto
Taxa de *Crossover*	98%
Taxa de mutação	0,1%

Vale ressaltar que o modelo de otimização utilizado não se restringe apenas à descoberta dos valores máximos de consumo, mas também pode ser utilizado para identificar os cenários que acarretam um consumo mínimo, médio ou um valor possível de

ser alcançado pela concessionária de energia, dada à variação dos aspectos econômicos considerados.

6.4.5. Descoberta do cenário sócio-econômico ótimo que explique a maximização do consumo residencial e comercial

Para mostrar o emprego da estratégia de descoberta de cenários, no qual mais de uma variável do domínio é considerada como meta, é considerada uma situação real e demandada pelo especialista, relacionada à descoberta do cenário sócio-econômico que mais esteja associado à maximização da variável consumo de energia elétrica residencial e consumo de energia elétrica comercial.

Para efeito de estudo de caso, são utilizadas as bases de dados da ENERSUL e do Governo do Estado do Mato Grosso do Sul (informações sócio-econômicas). A idéia básica é determinar o cenário sócio-econômico mais propício para a obtenção do valor máximo de consumo residencial e comercial, considerando duas análises:

- Variáveis metas consumo residencial e consumo comercial com os mesmos pesos; e

- Variáveis metas consumo residencial e consumo comercial com pesos 0,6 e 0,4, respectivamente, face à necessidade de se avaliar o maior impacto, demandado pelo especialista do domínio, do consumo residencial sobre o total faturado de energia elétrica.

Na primeira análise, o AG obteve o cenário representado na tabela 6.4, com a probabilidade máxima de 0,674054, para obtenção dos consumos máximos de energia residencial e comercial.

Tabela 6.4. Cenário mais provável para maximização dos consumos residencial e comercial da ENERSUL, ambas com pesos iguais a 0,5.

Variável	Estado
RURAL_MWh	7
RESIDENCIAL_MWh	M
ILUMINACAO_PUBLICA_MWh	7
PIB_mensal_RS_milhoes	5
ExpCereais_ton	7
ExpSementes_e_Frutos_Oleaginosos_Graos_ton	6
Arrecadacao_de_ICMS_mil	9
ExpGorduras_oleos_Vegetais_ton	6
PIB_mensal_USS_milhoes	7
ExpAcucares_ton	4
ExpMinerios_ton	4
ExpCouros_e_Peles_ton	8
ExpMadeira_ton	3
ExpFerro_e_Aco_ton	4
Transf_da_Uniao_mil	4
INDUSTRIAL_MWh	3
COMERCIAL_MWh	M
PODER_PUBLICO_MWh	1

Na segunda análise, o AG obteve o cenário representado na tabela 6.5, com a probabilidade máxima de 0,787243, para obtenção dos consumos máximos de energia residencial e comercial.

Tabela 6.5. Cenário mais provável para maximização dos consumos residencial e comercial da ENERSUL, com pesos iguais a 0,6 e 0,4, respectivamente.

Variável	Estado
RURAL_MWh	7
RESIDENCIAL_MWh	M
ILUMINACAO_PUBLICA_MWh	7
PIB_mensal_RS_milhoes	3
ExpCereais_ton	7
ExpSementes_e_Frutos_Oleaginosos_Graos_ton	7
Arrecadacao_de_ICMS_mil	2
ExpGorduras_oleos_Vegetais_ton	2
PIB_mensal_USS_milhoes	9
ExpAcucares_ton	8
ExpMinerios_ton	3
ExpCouros_e_Peles_ton	3
ExpMadeira_ton	2
ExpFerro_e_Aco_ton	7
Transf_da_Uniao_mil	4
INDUSTRIAL_MWh	9
COMERCIAL_MWh	M
PODER_PUBLICO_MWh	1

É possível observar que a probabilidade de se alcançar o consumo máximo de energia residencial e comercial é maior quando atribuímos um peso maior ao consumo

residencial, o que corrobora o conhecimento especialista a respeito do maior impacto do consumo residencial, em relação ao consumo comercial, sobre o total faturado de energia elétrica.

6.4.6. Descoberta do cenário climático ótimo que propicie um valor máximo do consumo total de energia, considerando o grau de interesse.

Para mostrar o emprego da estratégia de descoberta de cenários, desta feita considerando o conhecimento *a priori* do especialista sobre o grau de interesse do padrão (cenário) a ser descoberto, suponha a tarefa de busca de um cenário ótimo de um valor meta do consumo total de energia elétrica, estabelecido por um especialista de mercado de energia da ENERSUL, considerando o grau de interesse, medido a partir da crença α_1 desse especialista de que "o consumo de energia aumenta na medida em que o índice pluviométrico diminui e a temperatura máxima aumenta".

Assim, poderíamos executar a estratégia de descoberta de cenário considerando α_1, calculada, via método de inferência sobre a rede da figura 6.6, com base em *P(consumo = 10|ind_pluv = 1, temp_max = 10),* no qual *ind_pluv = 1*, representa a menor faixa de valores de índice pluviométrico (0 a 36,76 mm), *temp_max = 10,* identifica a maior faixa de valores para temperatura máxima (39,2 ºC a 40,8 ºC) e *consumo = 10,* a maior faixa de valores de consumo faturado de energia (301.854 a 311.087 MWh).

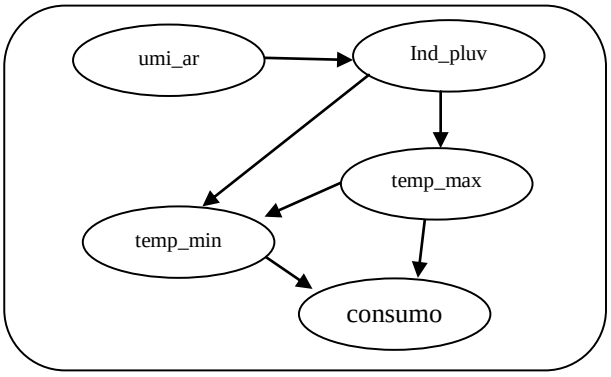

Figura 6.6. Rede bayesiana ENERSUL – fatores climáticos.

Desse modo, obtém-se o valor $P(\alpha_1|\xi)=P(consumo=10|ind\_pluv=1, emp\_max10)=0,57$. Esse valor é utilizado como entrada para o algoritmo

DESCOBERTA_CENÁRIO que, baseado na avaliação das aptidões dos indivíduos do AG, devolve o cenário ótimo que considera a crença inicial do usuário. O primeiro indivíduo (cenário) que atender ao critério de parada, estabelecido a partir do seu grau de interesse, será dado como resposta. Por exemplo, considerando o grau de interesse igual a *0,25*, o primeiro cenário encontrado pelo AG tem a seguinte configuração (tabela 6.6):

Tabela 6.6. Valores dos atributos para maximização do valor do consumo.

Atributo	**Valor**
ind_pluv	0 a 36,76 mm
umi_ar	56,2% a 67,4%
temp_max	39,2 ºC a 40,8 ºC
temp_min	7,9 ºC a 9,9 ºC

Essa abordagem é particularmente interessante para os especialistas do setor elétrico, principalmente para os que a utilizam para efetuar análise de cenários para comercialização de energia elétrica, posto que os especialistas, via de regra, valem-se não apenas dos métodos de previsão de consumo de energia, mas também de suas experiências e crenças sobre os impactos de determinadas variáveis exógenas sobre as variações do consumo de energia elétrica (e.g. variáveis climáticas e econômicas).

6.5. CONSIDERAÇÕES FINAIS

Neste capítulo, foi possível estabelecer a instanciação da estratégia de descoberta de cenários no âmbito do setor elétrico, especialmente relacionada ao consumo de energia elétrica e seu relacionamento com aspectos climáticos e sócio-econômicos.

A estratégia proposta é relevante no contexto do setor elétrico, visto que possibilita a quantificação do relacionamento causal entre as variáveis de econômicas, climáticas e de consumo, por meio da descoberta dos valores que compõem uma combinação ótima, com vistas a obter uma determinada meta, por exemplo, de consumo. As abordagens da estratégia proposta foram criadas, principalmente, a partir das demandas desse setor da economia.

Os mecanismos criados para aprimorar o processo de inferência das redes bayesianas são bastante úteis, principalmente nos processo de comercialização de energia, no qual previsões de consumo de energia e o estudo de cenário para apoiar essas previsões são determinantes para o êxito nos processos de compra e venda de energia.

O interesse dos envolvidos neste Projeto em empregar as abordagens do modelo para diversos outros cenários, não apenas relativos ao consumo de energia, mas também às ações de governo (e.g. descoberta das variáveis, além dos valores das mesmas, que maximizam a geração de emprego e renda), tem encorajado a utilização da estratégia proposta. Esse interesse é comprovado pela utilização dos resultados dos estudos realizados nesta tese, no sistema de suporte à decisão implementado para outros estados brasileiros (e.g. Mato Grosso, Mato grosso do Sul e Tocantins), atendidos pelo mesmo grupo de empresas do qual faz parte a concessionária de energia elétrica do Pará.

Por mais que os estudos aqui desenvolvidos sejam eminentemente aplicados sobre os dados das empresas do Grupo Rede Energia, é imperativo compreender que as abordagens da estratégia desenvolvida são de grande valia não só para o caso dessas empresas, mas também podem ser generalizadas para outras ações e soluções de problemas de sistemas elétricos de potência.

É importante ressaltar que o sucesso das soluções propostas para a realização de projetos neste importante setor da economia tem sido corroborado por um conjunto de projetos já realizados ou em curso pela equipe de Mineração de Dados do LPRAD e que, de certa forma, atestam a experiência desse grupo, bem como a efetividade dos métodos, técnicas e ferramentas implementadas pelos mesmos para a consecução de projetos em diversos domínios de aplicação. No Anexo II, são apresentadas algumas das principais contribuições, materializadas por meio de projetos de pesquisa e desenvolvimento, da equipe para o setor elétrico.

7. CONCLUSÕES E TRABALHOS FUTUROS

Um dos pontos almejados por concessionárias do setor elétrico é a capacidade de planejamento de aquisição/venda de energia elétrica em um tempo futuro, dada a grande variação que esse mercado está exposto. Via de regra, as companhias de distribuição de energia provisionam o consumo de energia baseado apenas nos valores históricos do próprio consumo. Entretanto, a precisão dos valores previstos é comprometida, principalmente em regiões em plena expansão de sua rede e muito suscetíveis às variações climáticas e/ou econômicas, como a Região Amazônica.

Dessa forma, uma ferramenta útil para as concessionárias deve estabelecer métricas para o impacto que outras variáveis aleatórias exercem sobre o consumo de energia, de tal sorte que seja possível prever cenários cujas configurações de operação do sistema elétrico sejam econômicos, seguros, confiáveis e de boa qualidade.

Nesse horizonte, são recorrentes, no setor elétrico, demandas relacionadas à necessidade de se estabelecer cenários que permitam quantificar o impacto que determinadas variáveis econômicas e climáticas têm sobre o consumo de energia. Além disso, nos cenários descobertos, apontar a influência dessas variáveis sobre o consumo de energia, bem como descobrir os valores (estados) das variáveis econômicas e climáticas que propiciam um valor meta de consumo de energia elétrica. Visando atender a essas demandas e, principalmente, estabelecer um mecanismo que possa generalizar o processo de interpretação analítica das correlações entre as variáveis de um domínio de aplicação qualquer, foi criada uma nova estratégia, baseada na combinação de duas técnicas amplamente utilizadas em Mineração de Dados – Redes Bayesianas e Algoritmos Genéticos.

Para isso, as redes bayesianas são geradas a partir dos dados históricos econômicos, climáticos e de consumo faturado de energia. Para efeito de estudo de caso, são utilizados os dados históricos de consumo de energia, providos pelas Companhias de Energia Elétrica do Estado do Pará e do Mato Grosso do Sul, os dados econômicos fornecidos pela Secretaria de Planejamento desses estados, além dos dados climáticos, fornecidos pelo Instituto Nacional de Pesquisas Espaciais. Uma vez descobertas as correlações, via RB, das variáveis que influenciam diretamente no consumo, pode-se estabelecer os cenários que permitem que os usuários de níveis decisórios possam prever quais condições econômicas (estados das variáveis econômicas) favorecem a obtenção de um valor meta de consumo de energia, estabelecido por esses usuários.

Neste contexto, o ponto central deste trabalho é estabelecer um conjunto de estratégias para estender as potencialidades das RBs, no que concerne aos processos de inferência, com o objetivo precípuo de oferecer ao usuário um modo analítico de observação do domínio, por meio da descoberta de cenários que melhor expliquem a obtenção de uma meta estabelecida.

As contribuições deste trabalho podem ser caracterizadas em três aspectos: o primeiro, relacionado ao aperfeiçoamento das técnicas de Mineração de Dados. O segundo, ao provimento de instrumentos sólidos de suporte à decisão, ao setor elétrico. Por último, as contribuições inerentes ao processo de construção de um trabalho científico desta natureza.

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

- Proposta de uma estratégia que permita estender o poder de interpretabilidade das RBs, quantificando o relacionamento causal entre as variáveis do domínio, por meio da descoberta dos valores que compõem uma combinação ótima de estados, aqui chamada de cenário, com vistas a obter uma determinada meta;
- A estratégia criada oferece um conjunto de abordagens, o que representa uma mecanismo bastante abrangente de análises sobre as RBs, envolvendo:
    - a facilidade de uso de métodos de inferência em redes bayesianas exatos e aproximados como função de avaliação dos cenários;

- a possibilidade de encontrar as variáveis que exercem maior influência na obtenção de um valor meta;
- a capacidade de se obter os valores contínuos das variáveis de evidência;
- a incorporação de conhecimento do especialista para conduzir o processo de busca da melhor configuração das variáveis de consulta;
- a descoberta de cenários (configuração mais provável) com mais de uma variável meta, ponderadas pelo especialista, no contexto do domínio de aplicação.

- Aplicar a estratégia e suas abordagens em situações associadas a problemas do mundo real;
- Utilização de AGs para melhoria da interpretabilidade, ao invés de utilizá-lo na otimização da estrutura e dos parâmetros da RB, como é comumente referenciado na literatura;
- Empregar técnicas híbridas de Inteligência Computacional com vistas à melhoria dos processos de descoberta de conhecimento em bases de dados;
- Permitir a análise de cenários antecipatórios em RBs, ao contrário de como essas redes são frequentemente utilizadas, para a obtenção de cenários exploratórios;
- Adequar ainda mais os sistemas de Suporte à Decisão baseados em MD às aplicações do mundo real, provendo esses sistemas de outros modos de interpretação e inferências;

No que concerne à aplicação das técnicas de Mineração de Dados no setor elétrico, temos a considerar:

- Estabelecimento de mecanismos que permitem quantificar a influência que determinadas variáveis econômicas e climáticas exercem sobre o consumo faturado de energia elétrica;
- Criação de uma estratégia para descoberta de uma combinação ótima dos estados das variáveis (climáticas e sócio-econômicas), com vistas a obter um determinado valor meta de consumo de energia elétrica. É possível destacar que

a estratégia não se restringe apenas a estabelecer cenários econômicos e climáticos que propiciem uma meta de consumo, mas também permite realizar análises das correlações entre os próprios dados de consumo (e.g. entre as classes de consumo), além de poder estabelecer, o contrário, ou seja, cenários de consumo de energia que melhor explique a obtenção de um valor meta de uma ou mais variáveis econômicas, ou até mesmo climáticas;

- Um efetivo referencial de análise para pautar decisões relacionadas ao processo de comercialização de energia, onde previsões de consumo de energia e o estudo de cenário para apoiar essas previsões são determinantes para o êxito nos processos de compra e venda;

- A partir da estratégia desenvolvida, foi possível criar um completo ambiente de suporte à decisão, já em operação, para as concessionárias de energia elétrica que atuam em seis estados brasileiros (Pará, Mato Grosso, Mato grosso do Sul, Tocantins, além de algumas regiões do sul de São Paulo e do Norte do Paraná), atendidas pelo Grupo Rede Energia, a fim de que os usuários de níveis decisórios dessas empresas possam estabelecer contratos mais vantajosos no mercado futuro de energia e analisem cenários de consumo favoráveis, baseados em variações climáticas e econômicas, para as suas operações no setor. Adicionalmente, as análises podem também estar relacionadas às ações dos governos desses estados, por exemplo, para descobrir cenários, envolvendo energia, clima e dados econômicos que maximizam a geração de emprego e renda.

Em observação às contribuições inerente ao processo de consecução de um projeto científico desta natureza, podem ser destacadas:

- Divulgação dos estudos realizados neste trabalho, junto às comunidades nacional e internacional, por meio da publicação de artigos em congressos e periódicos, o que corrobora a importância e contribuições das investigações realizadas. No Anexo I, são apresentadas as publicações realizadas no decorrer do desenvolvimento da tese e que possuem relação com as pesquisas que alicerçaram a mesma.

- Elaboração do documento de tese para disponibilização acadêmica dos estudos, métodos e aplicação desenvolvidos, bem como dos resultados obtidos.
- Desenvolvimento de pesquisas em Mineração de Dados, a partir de estudos como os que fundamentam essa tese, e que estão sendo aplicados nas soluções propostas nos projetos, cujos exemplos são demonstrados na tabela 6.7. Os resultados acadêmico-científicos dessas investigações e projetos encorajam o fortalecimento da massa crítica local com vistas à consolidação de um grupo de pesquisa especializada em Mineração de Dados para o Setor Elétrico.

Cabe ressaltar, que ao longo do desenvolvimento da tese, algumas dificuldades foram encontradas para elaboração e aplicação da estratégia desenvolvida. Dentre essas dificuldades, podem ser destacadas as que seguem:

- Como pode ser observada na literatura, uma das grandes dificuldades encontradas na execução de processo de Mineração de Dados, está relacionada à identificação das fontes e o pré-processamento dos dados. Neste trabalho não foi diferente. As bases de dados utilizadas estavam distribuídas e pertenciam a fontes diversas (dados oriundos do poder público estadual, de institutos de pesquisa e de empresas privadas). Além disso, foram necessárias análises sobre domínios econômicos e climáticos de diferentes estados da federação, o que aumentou o grau de dificuldade de compreensão dos mesmos.
- Compreensão do formalismo matemático próprio dos métodos que fundamentam as técnicas empregadas em Mineração de Dados.

A partir das contribuições e dificuldades aqui elencadas e pela própria natureza do processo de construção de um trabalho científico, há sempre margem para desdobramentos dos estudos executados, aprimoramentos e aplicações que transcendem ao escopo das investigações realizadas. Por essas razões, alguns temas podem ser apontados como sugestões de trabalhos futuros às pesquisas que fundamentam esta tese, os quais são pontuados a seguir.

Como há um interesse intrínseco do grupo do LPRAD por aplicações de MD no setor elétrico, algumas novas aplicações podem ser realizadas, via estratégia de descoberta de cenários:

- Analisar cenários, considerando fatores como a demanda de energia advinda de novos projetos de desenvolvimento dos estados pesquisados (e.g. novos empreendimentos na área de siderurgia e seus impactos econômico na região sudeste deste Estado), além de programas para atendimento ao setor rural, capitaneados pelo Programa "Luz para todos", do Governo Federal;
- Estabelecer cenários de otimização do consumo, considerando diversos outros fatores sugeridos pelo especialista do domínio, como segurança do sistema, qualidade da energia suprida aos consumidores, ou até mesmo outros elementos que norteiam o processo de comercialização de energia (e.g. descoberta de cenários que considerassem o PLD, de acordo com os níveis dos reservatórios e a expansão do sistema hidrotérmico);

Outra sugestão para aplicação da estratégia seria em situações em que o tempo fosse um fator determinante (e.g. controle de tráfego). Assim, uma possível utilização seria prever cenários que pudessem proporcionar uma determinada situação (evento) e fossem identificadas as configurações das variáveis do domínio que pudessem potencializar ou evitar esse evento. Face à condição crítica do tempo, a estratégia poderia ser modificada, por exemplo, considerando a paralelização do AG, com vistas ao aumento de seu desempenho.

Quanto aos desdobramentos da estratégia propriamente dita, podem ser elencadas as seguintes sugestões de continuidade:

- Investigação de outros métodos de otimização de modo a compará-lo em termos de desempenho e precisão com o algoritmo genético utilizado neste trabalho;
- Novos mecanismos para considerar o conhecimento *a priori* do especialista sobre o domínio, considerando outras medidas subjetivas (e.g. grau de utilidade);

As sugestões propostas de trabalho futuro podem, evidentemente, ser combinadas, de tal sorte que seja possível utilizar os desdobramentos da estratégia desenvolvida em aplicações no setor elétrico distintas das que foram apresentadas aqui, bem como em outros domínios de aplicação.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ABDELBAR, A.M.; HEDETNIEMI, S.M. A parallel hybrid genetic algorithm simulated annealing approach to finding most probable explanations on Bayesian belief networks. In: Proceedings IEEE International Conference on Neural Networks Vol. I, pp. 450–455 (1997).

ABRAMSON, B.; BROWN, J.; EDWARDS, W.; MURPHY, A.; WINKLER, R. L. Hailfinder: A Bayesian system for forecasting severe weather. International Journal of Forecasting, Elsevier, vol. 12(1), pages 57-71 (1996).

ANDERSEN, S. L.; OLSEN, K. G.; JENSEN, F. V.; JENSEN, F. Hugin – A shell for building Bayesian belief universes for expert system. In Proceeding of the 11th International Joint Conference on Artificial Intelligence. Spring-Verlag, (1): 10800085 (1989).

ANDREASSEN, V.; JENSEN, F.V.; ANDERSEN, S.K.; FALCK, B.; KJAERULFF, U.; WOLDBYE. M.; SORENSEN, A.; ROSENFALCK, A. & JENSEN, F. MUNIN – an expert EMG assistant. In: Computer-aided electromyography and expert systems. DESMEDT, J.E. (ed.). Amsterdam: Elsevier Science, p. 255-277 (1989).

BARROS, M.; MELLO, M.; SOUZA, R. Aquisição de energia no mercado cativo brasileiro: simulações dos efeitos da regulação sobre o risco das distribuidoras. Pesquisa Operacional, vol.29, n.2, pp. 303-322 (2009).

BEASLEY, D.; BULL, D.R.; MARTIN, R.R. "An overview of Genetic Algorithms: part 1, fundamentals". University Computing, 15(2): 58-69 (1993).

BEINLICH, I.; SUERMONDT, H.; CHAVEZ, R.; E COOPER, G. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In Proceedings of the Second European Conf. on Artificial Intelligence in Medicine, volume 38, páginas 247–256. (1989).

BUNTINE, W. A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 8(3): 111034 (1996).

CASELLA, G.; BERGER, R. L. Statistical inference. Wadsworth and Brooks (1990).

CASTILLO, E.; GUTIERREZ, J.; HADI; A. Expert systems and probabilistic network models. Springer (1997).

CHEN, Z. Computational intelligence for decision support. CRC PRESS (1999).

CHEN, Z. Data mining and uncertain reasoning - an Integrated Approach. John Wiley Professional (2001).

CHENG, J.; BELL, D. A.; LIU, W. Learning belief networks from data: an information theory based approach, Proceedings of the sixth international conference on Information and knowledge management, (1): 325-331 (1997).

CHENG. J.; DRUZDZEL, M. J. BN-AIS: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. Journal of Artificial Intelligence Research, 13:155.188 (2000).

CHIANG, C. Genetic Algorithm for Static Power Economic Dispatch. Computer Science and Information Engineering, World Congress on, vol. 4, pp. 646-650, 2009 WRI World Congress on Computer Science and Information Engineering (2009).

CHU, P. C.; BEASLEY J. E. A genetic algorithm for the generalized assignment problem. Comput. Operat. Res., 24 (1): 17–23 (1997).

COOPER, G.; HERSKOVITZ, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning, (9): 309–347 (1992).

DAGUM, P. LUBY, M. Approximating probabilistic inference in Bayesian belief networks is NPhard, Artificial Intelligence 60 141–153 (1993).

DAHAL, K. P.; ALDRIDGE, C. J.; GALLOWAY, S. J. Evolutionary hybrid approaches for generation scheduling in power systems. European Journal of Operational Research. Elsevier, 177(3): 2050-2068 (2007).

D'AMBROSIO, B. Inference in Bayesian networks, AI Magazine, 20(2), 21:36, (1999).

DAWID, P. Statistical theory – the frequential approach. Journal of the Royal Statistical Society. (147): 178-292 (1984).

DE CAMPOS, L.M.; GAMEZ, J.A.; MORAL, S. Partial abductive inference in Bayesian belief networks - an evolutionary computation approach by using problem-specific genetic operators. Evolutionary Computation. IEEE Transactions, 6 (2): 105–131 (2002).

DECHTER, R. RISH, I. Mini-buckets: A general scheme for bounded inference, Journal of the ACM (JACM), 50(2), 107:153 (2003).

DIETTERICH, T. G. Machine learning research: Four current directions. Artificial Intelligence Magazine, 18(4): 97036 (1997).

DILLON, W. R.; GOLDSTEIN, M: Multivariate analysis - methods and applications. John Wiley & Sons (1984).

ESHELMAN, L. J.; CARUANA, R. A.; SCHAFFER, J. D. Biases in the crossover landscape. Proceedings of the Third International Conference on Genetic Algorithms, Morgan Kaufmann (1): 1009 (1989).

FAYYAD, U. Data Mining and Knowledge Discovery: making sense out of data. IEEE Expert, 11(5): 20-25 (1996).

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. Communication of the ACM, 39 (11): 27-34 (1996).

FLORES-LOREDO, Z.; IBARGUENGOYTIA, P. H.; MORALES, E. F. On line diagnosis of gas turbines using probabilistic and qualitative reasoning. Intelligent Systems Application to Power Systems. Proceedings of the 13th International Conference on, (1): (2005).

FLORES-QUINTANILLA, J. L.; MORALES-MENENDEZ, R.; RAMIREZ-MENDOZA, R. A.; GARZA-CASTANON, L.E.; CANTU-ORTIZ, F. J. Towards a new fault diagnosis system for electric machines based on dynamic probabilistic models. Proceedings of the American Control Conference on, (4): 2775-2780 (2005).

FUNG, R.; CHANG. K.C. Weighing and integrating evidence for stochastic simulation in Bayesian networks. In M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, editors, Uncertainty in Artificial Intelligence 5, pages 209.219, New York, N. Y. Elsevier Science Publishing Company, Inc (1989).

GAAG, L. Bayesian belief networks: odds and ends. The Computer Journal, (39): 97–113 (1996).

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. ACM SIGKDD, (1): 20-33 (1999).

GOLDBERG, D.E. Genetic Algorithms in search, optimization and machine learning. Addison-Wesley (1989).

GOLDSCHMIDT, R.; PASSOS, E. Data Mining: um guia prático. Campus-Elsevier (2005).

GREFENSTETTE, J. Optimization of control parameters for genetic algorithms. IEEE Transactions on Systems, Man and Cybernetics. 16 (1): 122028 (1986).

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Multivariate data analysis. Prentice-Hall (1998).

HAN, J.; KAMBER. M. Data Mining: concepts and techniques. Morgan Kaufmann (2001).

HAN, J.; KAMBER. M. Data Mining: concepts and techniques. Morgan Kaufmann. 2ª edição (2006).

HANDA, H.; KATAI, O. Estimation of Bayesian network algorithm with GA searching for better network structure. Neural Networks and Signal Processing, Proceedings of the International Conference on, (1): 436–439 (2003).

HAYKIN, S. Redes Neurais – Princípios e Prática. Bookman (2001).

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning, 20: 197–243 (1995).

HECKERMAN, D. Bayesian networks for Data Mining. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, (1): 79019 (1997).

KJAERUFF, U. A computational scheme for reasoning in dynamic probabilistic networks. In Uncentainty in Artificial Intelligence: Procedings of the Eighth Conference on. Morgan Kaufmann, (1): 121029 (1992).

KORB, K. B.; NICHOLSON, A. E. Bayesian Artificial Intelligence. Chapman & Hall/CRC (2003).

KRISTENSEN, K.; RASMUSSEN, I.A.. Models to assist the choice of strategies for growing malting barley without the use of pesticides. In: Secher, B.J.M.; Rossi, V. & Battilani, P. (eds.): Workshop on Computer-based DSS on Crop Protection. SP-report no. 7, 39-46 (1993).

KUMAR, P.; CHANDNA, V. K.; THOMAS, M. S. Fuzzy-Genetic Algorithm for pre-processing data at the RTU. IEEE Transactions on Power Systems, 19 (2): 718-723 (2004).

KUO, B.; HSIEH, T.; Wang, H. Using MPE with Bayesian Network for Sub-optimization to Entropy-Based Methodology. In Proceedings of the Eighth international Conference on intelligent Systems Design and Applications - Volume 01. IEEE Computer Society, 381-386 (2008).

LAM, W.; BACCHUS, F. Learning bayesian belief networks: An approach based on the MDL principle. Computational Intelligence, 10(4): 269–293 (1994).

LAURITZEN, S. L.; SPIEGELHALTER, D. J. Local computations with probabilities on graphical structures and their application to expert systems. J. Royal Statistics Society B, 50, 157–194 (1988).

LIU, C.-L.; WELLMAN, M. P., Evaluation of Bayesian networks with flexible state-space abstraction methods, International Journal of Approximate Reasoning 30(1), 1–39 (2002).

LOPES, C. H. P.; PACHECO, M.A.C.; VELLASCO, M. M. B. R.; PASSOS, E. P. L. RULE-EVOLVER: An evolutionary approach for Data Mining. Proceedings of The Seventh International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Lecture Notes in Artificial Intelligence 1711, New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, (1): 458-462 (1999).

LUNA, J. E. O. Algoritmos EM para aprendizagem de redes Bayesianas a partir de dados incompletos. Dissertação de Mestrado, Departamento de Computação e Estatística - Centro de Ciências Exatas e Tecnologia - Universidade Federal de Mato Grosso do Sul (2004).

MALACHI, Y; SINGER, S. A genetic algorithm for the corrective control of voltage and reactive power. IEEE Transactions on Power Systems, 21 (1): 295-300 (2006).

MALHAS, R.; AGHBARI, Z. A. Interestingness filtering engine: Mining Bayesian networks for interesting patterns. Expert Systems with Applications 36, 3: 5137-5145 (2009).

MENGSHOEL, O.; WILKINS, D.; ROTH, D. Controlled generation of hard and easy Bayesian networks: Impact on maximal clique size in tree clustering, Artificial Intelligence, v.170 n.16-17, p.1137-1174 (2006).

MICHALEWICZ, Z. Genetic algorithms + data structures = evolution programs. AI Series. Springer-Verlag (1994).

MORALES, M. M.; DOMINGUEZ, R. G.; RAMIREZ, N. C.; HERNANDEZ, A. G.; ANDRADE, J. L. J. A method based on genetic algorithms and fuzzy logic to induce Bayesian networks. IEEE Proceedings of the Fifth Mexican International Conference in Computer Science, (1): 176080 (2004).

MISHRA, S.; REDELY, G.D.; RAO, P.E.; SANTOSH, K. Implementation of new evolutionary techniques for transmission loss reduction. In Proc. IEEE Congress on Evolutionary Computation CEC 2007, pages 2331–2336, 25–28 (2007).

NEAPOLITAN, R.E. Probabilistic Reasoning in Expert Systems: Theory and Algorithms, Wiley, New York (1990).

NEGOITA, M. G. Book review, Data Mining methods for Knowledge Discovery, SIGKDD, 1 (2): 118019 (2000).

NICHOLSON, A.E.; JITNAH, N. Belief network algorithms: A study of performance using domain characterization, in: G. Antoniou, A.K. Ghose, M. Truszczynski (Eds.), Learning and Reasoning with Complex Representations, Lecture Notes in Artificial Intelligence, vol. 1359, Springer, Berlin, pp. 169–188 (1998).

PAVÓN, R.; DÍAZ, F.; LAZA, R.; LUZÓN, V. Automatic parameter tuning with a Bayesian case-based reasoning system. A case of study. Expert Syst. Appl. 36, 2, 3407-3420 (2009).

PEARL, J. Probabilistic reasoning in Intelligent System, Morgan Kaufmann Publishers (1988).

PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. The Interestingness of Deviations. Proc AAAI '94 Workshop Knowledge Discovery in Databases, pp 25-36 (1994).

QUINLAN, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993).

RAMONI, M.; SEBASTIANI, P. Discovering Bayesian networks in incomplete databases. Knowledge Media Institute, The Open University, Kmi Technical Report, nº 46 (1997).

RAMONI, M.. Robust learning with missing data. Machine Learning, 45(2): 147–170 (2001).

RAMOS, F.T.; COZMAN, F.G. Anytime anyspace probabilistic inference, International Journal of Approximate Reasoning, 38, 53:80 (2005).

REIZ, B.; CSATÓ, .; DUMITRESCU, D. Prüfer Number Encoding for Genetic Bayesian Network Structure Learning Algorithm. Symbolic and Numeric Algorithms for Scientific Computing, International Symposium on, pp. 239-242. 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (2008).

REZENDE, S. O. Sistemas Inteligentes – Fundamentos e Aplicações. Manole (2003).

ROCHA, C. A. J. da. Redes Bayesianas para extração de conhecimento de bases de dados, considerando a incorporação de conhecimento de fundo e o tratamento de dados incompletos. Dissertação de Mestrado, ICMC-USP (1999).

RUSSEL, S.; NORVIG, P. Artificial Intelligence – a modern approach. Prentice Hall (2003).

SALZBERG, S. L. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery, 1 (3): 317-328 (1997).

SANTANA, A. L.; ROCHA, C. A. J. da; FRANCÊS, C. R. L.; RÊGO, L. P.; VIJAYKUMAR, N. L.; CARVALHO, S. V. de; COSTA, J. C. W. Strategies for improving the modeling and interpretability of Bayesian networks. Data & knowledge engineering, (63): 91007 (2007).

SANTANA, A. L. Estratégias para a melhoria da modelagem e interpretabilidade de redes bayesianas. Tese de Doutorado. Programa de Pós-Graduação em Engenharia Elétrica - Instituto de Tecnologia - Universidade Federal do Pará. Belém (2008).

SHIMONY, S.; DOMSHLAK, C. Complexity of probabilistic reasoning in directed-path singly-connected Bayes networks. Artificial Intelligence, v.151 n.1-2, p.213-225 (2003).

SILBERSCHATZ, A.; TUZHILIN, A. What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Transaction on Knowledge and Data Engineering 8, 6. 970-974 (1996).

SILVA, I. de J.; RIDER, M. J.; ROMERO, R.; MURARI, C. A. F. Transmission network expansion planning considering uncertainty in Demand. IEEE Transactions on Power Systems, 21(4): 15650573 (2006).

SOUZA, S. A. de; MACÊDO, R. A. de; VARGAS, E. T.; COURY, D. V.; OLESKOVICZ, M. Parameter estimation for an electric power system using genetic algorithms. IEEE Latin America Transaction. (4): 47-54 (2006).

SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. Causation, prediction, and search. MIT Press (2001).

STANKOVSKI V., SWAIN M., KRAVTSOV V., NIESSEN T., WEGENER D., KINDERMANN J., DUBITZKY W. Grid-enabling data mining applications with DataMiningGrid: An architectural perspective. Future Generation Computer Systems, 24 (4), pp. 259-279 (2008).

TAN, P.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. Addison Wesley (2005).

TURBAN, E.; ARONSON, J. E. Decision Support Systems and Intelligent Systems. Prentice-Hall (2001).

VERMA, D.; RAO, R. P. N.: Planning and Acting in Uncertain Environments using Probabilistic Inference. IEEE. 2382-2387 (2006).

WILSONA, A.G.; HUZURBAZAR, A.V. Bayesian networks for multilevel system reliability. Reliability Engineering & System Safety, 92,(10), pp.1413-1420. (2007).

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1, 67-82 (1997).

YANG, C. C. Fuzzy Bayesian inference. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 2707-2712 (1997).

YANG, G.Y.; DONG, Z.Y.; WONG, K.P. A modified differential evolution algorithm with fitness sharing for power system planning. IEEE Transactions on Power Systems, 23 2: 514-522 (2008).

YONGLI, Z.; LIMIN, H.; JINLING, L. Bayesian networks-based approach for power systems fault diagnosis. IEEE Transactions on Power Delivery, (21) 2: 634-639 (2006).

YONGLI, Z.; LIMIN, H. LIGUO, Z.; YAN, W. Bayesian Network Based Time-Sequence Simulation for Power System Reliability Assessment. *Mexican International Conference on Artificial Intelligence*, pp. 271-277. Seventh Mexican International Conference on Artificial Intelligence (2008).

YONGGIANG, W.; FANGCHENG, L.; HEMING, L. The fault diagnosis method for electrical equipment based on Bayesian network. Electrical Machines and Systems, (ICEMS). Proceedings of the Eighth International Conference on Publication, (3): 2259-2261 (2005).

ZHOU, Y.; PAHWA, A.; YANG, S. S. Modeling weather-related failures of overhead distribution lines. Power Systems, IEEE Transactions, 21(4): 1683 – 1690 (2006).

ANEXO I - TRABALHOS ACEITOS/PUBLICADOS

São apresentadas a seguir as referências aos principais trabalhos aceitos/publicados, que guardam relação com esta tese.

Capítulo de Livro:

ROCHA, Cláudio; CONDE, Guilherme; SANTANA, Ádamo; FRANCÊS, Carlos Renato; Rego, Liviane. Scenario discovery and temporal analysis for energy consumption forecasting of the Brazilian Amazon power suppliers. Nova Science Publishers (2009). (capítulo de livro aceito). Previsão de publicação: março/2010.

Periódicos:

1. SANTANA, Á. L. ; REGO, Liviane ; CONDE, G. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio ; SILVA, Marcelino S. da . Comparative Analyses of Computational Intelligence Models for Load Forecasting: a Case Study in the Brazilian Amazon Power Suppliers. Lecture Notes in Computer Science, v. 5553, p. 10440053 (2009).

2. REGO, Liviane ; SANTANA, Á. L. ; CONDE, G. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio . Predict - Sistema de Suporte à Decisão para Estimação de Cargas e Modelagem de Dependência em Sistemas Elétricos. Revista Pesquisa e Desenvolvimento da ANEEL, v. 3, p. 97000 (2009).

3. SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio ; REGO, Liviane ; VIJAYKUMAR, Nandamudi L. ; CARVALHO, Solon V. ; COSTA, João C. W. A. . Strategies for Improving the Modeling and Interpretability of Bayesian Networks. Data & Knowledge Engineering, v. 63, p. 91007 (2007).

4. CONDE, G. ; SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; REGO, Liviane ; ROCHA, Cláudio ; CARDOSO, Diego L. ; COSTA, João C. W. A. ; GATO, Vanja . Performance Evaluation of Short and Long Term Load Forecasting Models: a Case Study in the Amazonian Power Suppliers. Proceedings of SPIE, v. 6763, p. 67630V (2007).

5. SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio ; REGO, Liviane ; COSTA, João ; GATO, Vanja ; TUPIASSU, Armando . Decision Support in Power Systems Based on Load Forecasting Models and Influence Analysis of Climatic and Socio-Economic Factors. Proceedings of SPIE, the International Society for Optical Engineering, v. 6383, p. 63830I (2006).

6. SANTANA, Á. L. ; ROCHA, Cláudio; FRANCÊS, Carlos Renato Lisboa ; COSTA, João ; VIJAYKUMAR, Nandamudi L. ; CARVALHO, Solon V. . Strategies for improving the interpretability of Bayesian networks using Markovian time models and genetic algorithms. Proceedings of SPIE, the International Society for Optical Engineering, v. 6383, p. 63830R (2006).

7. ROCHA, Cláudio; SANTANA, A. L.; FRANCÊS, C. R. L.; FAVERO, E.; REGO, L., BEZERRA, U. H.; COSTA, J. C. W. A. Load forecasting and learning of influence patterns of the socio-economic and climatic factors on the power consumption. Wseas Transactions On Mathematics, 4(3): 176–184 (2005).

Trabalhos completos publicados em anais de congressos:

1. CONDE, G. ; SANTANA, Á. L. ; REGO, Liviane ; FRANCÊS, Carlos Renato Lisboa; ROCHA, Cláudio. Comparative Analyses of Computational Intelligence Models for Load Forecasting: a Case Study in the Brazilian Amazon Power Suppliers. In: International Symposium on Neural Networks (ISNN 2009), 2009, Wuhan. Sixth International Symposium on Neural Networks, 2009.

2. SANTANA, Á. L. ; REGO, Liviane ; CONDE, G. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio. Modelos de Inteligência Computacional Aplicados em um Estudo de Caso de Concessionária de Energia da Amazônia Brasileira para Previsão de Energia Elétrica. In: V CITENEL, 2009, Belém. Anais do V CITENEL, 2009.

3. ROCHA, Cláudio ; SANTANA, Á. L.; FRANCÊS, Carlos Renato Lisboa ; FAVERO, Eloi . Aplicação Híbrida de Algoritmos Genéticos para a Otimização e Melhoria da Interpretabilidade de Redes Bayesianas: uma Aplicação em Sistemas Elétricos de Potência. In: XXXIX Simpósio Brasileiro de Pesquisa Operacional, 2007, Fortaleza. Anais do XXXIX Simpósio Brasileiro de Pesquisa Operacional, 2007.

4. CONDE, G. ; SANTANA, Á. L. ; ROCHA, Cláudio; FRANCÊS, Carlos Renato Lisboa ; REGO, Liviane ; GATO, Vanja . Estratégias de Previsão de Carga e de Consumo de Energia Elétrica Baseadas em Modelos Estatísticos e Redes Neurais Artificiais: um Estudo de Caso nas Concessionárias de Energia do Estado do Pará. In: VIII Congresso Brasileiro de Redes Neurais (CBRN2007), 2007, Florianópolis. Anais do VIII Congresso Brasileiro de Redes Neurais (CBRN2007), 2007.

5. CONDE, G. ; SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; REGO, Liviane ; ROCHA, Cláudio ; GATO, Vanja . Comparative studies of Statistical and Neural Networks Models for Short and Long Term Load Forecasting: a Case Study in the Brazilian Amazon Power Suppliers. In: AI-2007 Twenty-seventh SGAI International Conference on Artificial Intelligence, 2007, Cambridge. Proceedings of The AI-2007 Twenty-seventh SGAI International Conference on Artificial Intelligence, 2007.

6. ROCHA, Cláudio; SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; REGO, Liviane; FAVERO, Eloi ; GATO, Vanja . Sistema de Suporte à Decisão para Predição de Cargas e Modelagem de Dependência em Sistemas Elétricos de Potência. In: XXXIII SEMISH - Seminário Integrado de Software e Hardware, 2006, Campo Grande, 2006.

7. ROCHA, Cláudio; SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa; BEZERRA, Ubiratan ; TUPIASSU, Armando ; GATO, Vanja . Sistema de Suporte à Decisão para Predição de Cargas e Análise de Influência dos Fatores Climáticos e Sócio-Econômicos em Sistemas Elétricos de Potência. In: SENDI - XVII Seminário Nacional de Distribuição de Energia Elétrica, 2006, Belo Horizonte - MG, 2006.

8. ROCHA, Cláudio; SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa; FAVERO, Eloi ; MACEDO, Valquiria ; BEZERRA, Ubiratan ; TUPIASSU, Armando ; GATO, Vanja ; REGO, Liviane ; COSTA, Rafael ; NASCIMENTO, Cibelle . Decision Support System for Power Systems Applying Load Forecasting and the Learning of Causal Relationships. In: Mathematical models and methods in applied sciences, 2005, Sofia, Bulgaria, 2005.

9. SANTANA, Á. L. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio ; TUPIASSU, Armando ; REGO, Liviane ; BEZERRA, Ubiratan . Sistema de Suporte à Decisão para Gerenciamento da Energia em Consumidores do Grupo A. In: III Congresso de Inovação Tecnológica em Energia Elétrica (III CITENEL), 2005, Florianópolis. Anais do III Congresso de Inovação Tecnológica em Energia Elétrica (III CITENEL), 2005. v. 1. p. 1-6.

Resumos expandidos publicados em anais de congressos:

SANTANA, Á. L. ; CONDE, G. ; FRANCÊS, Carlos Renato Lisboa ; ROCHA, Cláudio ; REGO, Liviane ; GATO, Vanja . Estratégias para Previsão do Consumo de Energia Elétrica através de Modelos Estatísticos e de Inteligência Computacional: Um Estudo de Caso nas Concessionárias de Energia do Estado do Pará. In: XXXIX Simpósio Brasileiro de Pesquisa Operacional, 2007, Fortaleza. Anais do XXXIX Simpósio Brasileiro de Pesquisa Operacional, 2007.

ANEXO II – PROJETOS DE PESQUISA E DESENVOLVIMENTO

São apresentados a seguir os principais projetos desenvolvidos pela equipe de Mineração de Dados do LPRAD, na área de Mineração de Dados, especificamente para o setor elétrico.

Projeto	Principal Contribuição	Financiador	Período
Estratégias de Melhoria da Interpretabilidade das Inferências Bayesianas Utilizadas no Sistema Predict	- Investigar a adequação da aplicação das técnicas de predição e de modelagem de dependência, utilizadas no Predict, nos dados relativos ao consumo e as condições sócio econômicas e climáticas das demais empresas do Grupo Rede: CEMAT, CELTINS e REDE SUL. - Implementar técnicas híbridas, utilizando Algoritmos Genéticos, Cadeias de Markov e Redes Bayesianas, objetivando estimar valores de consumo futuro a partir da correlação no tempo das variáveis de consumo e os fatores climáticos e sócio-econômicos; - Obtenção de valores de predição de energia associado a cenários alternativos das variáveis correlacionadas para planejamento de mercado, considerando, ainda, o potencial de economia de energia (energia conservada a ser investigada e caracterizada).	Agência Nacional de Energia Elétrica / Rede CELPA	2007-2009
PREDICT - Ferramenta de Suporte à Decisão para Predição de Cargas de Sistemas Elétricos	Projetar e implementar um sistema de suporte à decisão, utilizando métodos matemáticos e de inteligência computacional, para prever as necessidades de compra de energia no mercado futuro e para realizar inferências sobre a situação do sistema elétrico, a partir de dados históricos de consumo e suas correlações com aspectos sócio-econômicos e climáticos.	Agência Nacional de Energia Elétrica / Rede CELPA	2004-2006

Desenvolvimento de Tecnologias de Aquisição e Tratamento da Informação para Uso como Ferramenta de Aperfeiçoamento da Gerencia e Apoio à Decisão em Sistemas de Supervisão de Sistemas Elétricos de Potência que operam na Amazônia.	Estabelecer competência no estudo de estratégias de planejamento e apoio à decisão de políticas de desenvolvimento sustentável na Região Amazônica, a partir das informações privilegiadas que poderão ser adquiridas no processo de descoberta do conhecimento. Este objetivo neste projeto é específico: distribuição de energia. Entretanto, o arcabouço proposto é genérico e pode ser aplicado às várias possibilidades de ecossistemas, fato que norteia todo o projeto.	CNPq	2004-2006
Desenvolvimento de um Sistema Digital para Gerenciamento da Energia em Consumidores do Grupo A	Desenvolvimento de um Sistema de Suporte à Decisão, baseado em diversos modelos de inteligência computacional (tais como Bayesianas e SVM), para traçar o perfil de consumidores do Grupo A, atendidos pela Rede Celpa.	Agência Nacional de Energia Elétrica / Rede CELPA	2004-2005
Desenvolvimento de Estimador de Estados Robusto Utilizando Nova Técnica de Processamento de Erros	Desenvolver um estimador de estados generalizado capaz de identificar erros de topologia e de processar erros grosseiros em medidas, para obtenção da condição de operação real do sistema, bem como otimizar o estimador desenvolvido para a rede básica do Sistema Eletronorte, com a utilização de modelos adequados para a representação dos elementos da rede e do sistema de medição empregado na Empresa.	Eletronorte	2003-2004