UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS EM REDES DE BAIXA TENSÃO USANDO ALGORITMOS GENÉTICOS

FLORIANÓPOLIS
2005

ISABELA ANCIUTTI

MINERAÇÃO DE DADOS EM REDES DE BAIXA TENSÃO USANDO ALGORITMOS GENÉTICOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação pela Universidade Federal de Santa Catarina.

Orientador: Prof. Frank Augusto Siqueira.

FLORIANÓPOLIS
2005

ISABELA ANCIUTTI

MINERAÇÃO DE DADOS EM REDES DE BAIXA TENSÃO USANDO ALGORITMOS GENÉTICOS

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração Sistemas de Conhecimento, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Raul S. Wazlawick, Dr.

Banca Examinadora:

_____ Frank A. Siqueira, Dr.	_____ José L. Todesco, Dr.
_____ Paulo S. S Borges, Dr.	_____ Aran B. T. Morales, Dr.

Florianópolis / 2005

Dedico este trabalho a todos aqueles que buscam incessantemente o conhecimento, que constroem idéias, sonham com o infinito e se entregam corajosamente ao desafio de desvendar a si mesmos e ao mundo.

AGRADECIMENTOS

Em primeiro lugar a Deus, cuja graça e misericórdia se fizeram presentes durante todo o caminho. Toda a glória ao Senhor dos Exércitos, ao Rei dos Reis.

Ao meu orientador, Professor Frank, que me deu a chance de realizar esse sonho permitindo-me buscar um ideal mais elevado. Seus conselhos, orientação e compreensão foram aliados imprescindíveis durante todo o curso. Também aos professores da banca, por aceitarem conhecer e avaliar este trabalho. Ao Dr. Wesley Romão, por gentilmente ceder o código fonte do protótipo AGD, contribuindo de forma essencial para este trabalho.

À companhia CELESC, por permitir a utilização de seus dados neste estudo, além do acesso aos seus sistemas. Aos engenheiros e especialistas da empresa, que se dispuseram a colaborar, reconhecendo a importância da pesquisa. Em especial agradeço ao Eng. Marcelo Fernandes, ao Eng. Renato B. Rolim e ao Eng. Ricardo H. Guembarovski.

Ao Instituto Stela, que cedeu espaço para que este estudo fosse desenvolvido dentro de um de seus projetos. Obrigada principalmente a Isabel, cuja amizade e dedicação me são muito preciosas. Ao Marcio Napoli, pelas longas, pacientes e divertidas horas de apoio técnico. Grata a todos os colegas de trabalho e coordenadores, que muito me ensinaram.

A minha avó Yedda, pelo carinho e incentivo constantes. Mesmo estando em outro lado do País, o tempo todo esteve comigo, sendo acima de tudo uma amiga querida e um exemplo de bondade e generosidade. Ao meu tio Cesar, pelo apoio e por acreditar.

Aos meus amigos, que não apenas acreditaram em mim mas, principalmente, ajudaram-me a acreditar em mim mesma. Especialmente agradeço ao amigo distante Marcelo V. de Paula, mestre e aprendiz, para quem os horizontes são infinitos e a esperança é eterna. Não desista!

Ao meu amigo e amado Michael, que sempre será a estrela mais bela e brilhante no céu da minha vida. Um anjo para amar com toda a minha alma, um homem para amar com todo o meu coração.

SUMÁRIO

LISTA DE REDUÇÕES

AG	Algoritmo Genético
ADN	Ácido Desoxirribonucléico
ANEEL	Agência Nacional de Energia Elétrica
ARM	*Association Rule Mining*
BD	Banco de Dados
CELESC	Centrais Elétricas de Santa Catarina S.A.
CIF	Corporate Information Factory
COPEL	Companhia Paranaense de Energia
DEC	Duração Equivalente de Interrupção por Unidade Consumidora
DIC	Duração de Interrupção Individual por Unidade Consumidora
DM	*Data Mining*
DMIC	Duração Máxima de Interrupção Contínua por Unidade Consumidora
DRP	Duração Relativa da Transgressão de Tensão Precária
DRC	Duração Relativa da Transgressão de Tensão Crítica
DSS	Decision Support System
DW	*Data Warehouse*
EW	*Exploration Warehouse*
FEC	Freqüência Equivalente de Interrupção por Unidade Consumidora
FIC	Freqüência de Interrupção Individual por Unidade Consumidora
FP	Função de Pertinência
GENESIS	Gerência Integrada de Sistemas de Distribuição de Energia Elétrica
KDD	*Knowledge Discovery in Databases*
IA	Inteligência Artificial
IG	Impressão Geral
MBR	*Memory-Based Reasoning*
ODS	*Organization Decision Support*
OLAP	*Online Analytical Process*
PR	*Pattern Recognition*
RNA	Rede Neural Artificial
SA	Similaridade do Antecedente

SGBD	Sistema Gerenciador de Banco de Dados
SIMO	Sistema Integrado de Manutenção e Operação
TI	Tecnologia de Informação
UC	Unidade Consumidora
VHF	*Very High Frequency*
WEKA	Waikato Environment for Knowledge Analysis

LISTA DE FIGURAS

LISTAS DE TABELAS

LISTAS DE QUADROS

RESUMO

Diversos problemas atingem as redes de distribuição de energia elétrica no País. Entre eles, verificam-se aspectos fora de total administração e outros que podem ser previstos e posteriormente gerenciados e otimizados. Para limitar os efeitos destes problemas e incentivar a descoberta de soluções, a Agência Nacional de Energia Elétrica (ANEEL) definiu indicadores de confiabilidade que devem ser cumpridos pelas empresas no setor de distribuição elétrica, o que representou um significativo fator motivador na melhora dos serviços que prestam.

Nesse contexto, este trabalho propõe-se a auxiliar na previsão de falhas e otimização de problemas em redes de distribuição, extraindo conhecimento através da mineração de dados utilizando algoritmos genéticos em uma área relativamente nova no uso de tecnologia de informação para suporte às estratégias operacionais – o setor de energia elétrica no Brasil. Através da descoberta de regras de classificação, busca-se fornecer aos especialistas da CELESC – Centrais Elétricas de Santa Catarina – meios de incrementar os indicadores de confiabilidade da distribuição, permitindo a redução dos prejuízos causados pela interrupção do fornecimento de energia e melhorar a qualidade do serviço prestado a seus clientes.

Utilizando-se um *data warehouse* como fonte de dados e a experiência dos engenheiros especialistas, uma amostragem de dados foi processada e transformada. Uma ferramenta genético-difusa foi selecionada e adaptada ao ambiente do problema. A partir de três principais assuntos levantados pelos especialistas, o algoritmo genético foi executado e selecionaram-se regras de classificação conforme o seu *fitness* (calculado sobre a qualidade da regra e o seu grau de interesse – ambos envolvendo os dados referentes à freqüência relativa, cobertura e taxa de acerto da regra).

Os experimentos, considerando apenas 10% dos casos abrangidos pelas regras de classificação encontradas pelo algoritmo genético, estimaram que a companhia elétrica estudada deixou de arrecadar anualmente uma receita significativa devido à energia não-distribuída durante o período das interrupções. As regras de classificação extraídas, sua validade, simplicidade para a compreensão, utilidade prática, relevância no escopo do problema e interesse que representam aos analistas demonstraram a eficácia e o potencial da técnica de mineração de dados realizada neste estudo e aliada à experiência dos especialistas para extrair conhecimento do ambiente informacional de redes de distribuição de baixa tensão.

ABSTRACT

Several problems reach the electric power distribution in the Country. Among them, some aspects are found to be out of total administration, while some others can be foreseen and afterwards managed and optimized. To limit the effects of these problems and to encourage the solutions' discovery, the Agência Nacional de Elergia Elétrica (ANEEL) defined reliability indicators that should be reached by the companies in the sector of power distribution. This represented a significant factor to estimulate the improvement of the services which the companies provide.

In this context, this work aims to help in the failure issues and optimization prediction in the field of power distribution, extracting knowledge through data maning using genetic algorithms in an area relatively new on the use of information technology (IT) for support of operational strategies – the electric power sector in Brazil. Through the discovery of classification rules, this research aims to supply the specialists of CELESC – Santa Catarina's Electric Centrals – means to increase the reliability indicators of the power distribution, thus allowing the reduction of the damages caused by the power supply interruption and the improvement on the quality of the provided services to their customers.

Using a *data warehouse* as data source and the engineers' specialist experience, a data sample was processed and transformed. A genetic-diffuse tool was selected and adapted to the environment of the subject. Starting from three main subjects pointed by the specialists, the genetic algorithm was executed. Some classification rules were found according to their *fitness* (gathering quality of the rule and its interest level – both involving data about the relative frequency, coverage and hit rate of the rule).

The experiments, considering just 10% of the cases regarded by the classification rules found by the genetic algorithm, estimated that the studied electric company stopped levying annually significant revenue due to the electric power not distributed during the supply interruptions. The extracted classification rules, your legitimacy, comprehension simplicity, practical utility, relevance in the scope of the subject and the interest which they represent to the analysts, demonstrated the effectiveness and the potential of the *data mining* technique allied to the specialists' experience, in order to extract knowledge from the information environment of low voltage power distribuition.

1 INTRODUÇÃO

Todos os dias uma enorme quantidade de informações sobre as mais variadas áreas de conhecimento no mundo todo é recolhida e armazenada em meio digital. Atualmente se calcula que apenas 1% dessa informação esteja disponível na Web no formato de páginas virtuais (SUPER INTERESSANTE, 2004), enquanto a restante massa de dados coletados encontra-se nos sistemas corporativos, científicos e de domínio governamental.

Seguindo a filosofia de que todo conhecimento é poder, nunca antes na história as pessoas geraram e armazenaram tantos dados, principalmente nos setores financeiro e comercial. A abrangência do mercado, as tendências para investimento, a administração dos recursos disponíveis, o diferencial competitivo, entre outros aspectos, representam o grande fator motivador para saber não apenas mais, mas também antes e com segurança. A conclusão a que se chega é que na realidade a busca não é mais por informação comum, como foi algumas décadas atrás, porém, por um nível maior e mais raro dela: o conhecimento não trivial e aplicável.

Assim, o alvo mais simples de todo aquele que detém a informação ainda se mantém indubitável e persistente: alcançar conhecimento. Para a concretização desse objetivo os atuais sistemas de já são utilizados muito além da básica função de organização e padronização de dados ou do mero armazenamento e da disponibilização descritiva de conteúdo.

Do mesmo modo, como os conceitos de “informação” e “conhecimento” foram revistos com relação às fronteiras de suas definições (SCHREIBER, 2000) durante as pesquisas na área de sistemas de apoio à tomada de decisão, a estrutura de SGBDs[1] também foi obrigada a evoluir paralelamente, testando novas idéias de modelagem mais adequadas aos diferentes tipos de consulta dos usuários, distinguindo-se entre si quanto à funcionalidade dentro da organização, adaptando-se à tecnologia de hardware disponível para melhorar o desempenho, absorvendo conceitos de segurança, robustez, confiabilidade, integrando-se ou se tornando

[1] Sistema Gerenciador de Banco de Dados.

distribuída, conforme a disposição física requerida pela organização, o tipo de usuário consumidor, etc.

Em um nível superior da Tecnologia de Informação (TI), com o objetivo de explorar a informação e dela extrair conhecimento valioso e relevante, surgiu o conceito de Mineração de Dados (*Data Mining*), o qual tem sido constantemente aprimorado, discutido, abordado e aplicado em muitas áreas e setores da informação com o objetivo de entender, analisar e fazer uso dos dados (HUANG & WU, 2002). Em resumo, *Data Mining* (DM) pode ser definida como um conjunto de técnicas e ferramentas aplicadas à descoberta do conhecimento em bases de dados (ROMÃO et al., 2002).

O uso do termo “mineração de dados” deve-se à comparação bastante comum que se faz entre o potencial do imenso e oculto conhecimento inexplorado e um recurso mineral bruto na natureza que permanece encoberto sob uma camada sem valor de outros elementos. No tocante a esse recurso, as ferramentas e técnicas de extração de dados ainda são parte de uma tecnologia em constante desenvolvimento (MATHEUS et al., 1993), com grande crescimento quanto ao interesse e intensificação dos esforços em pesquisas – o que se comprova pelo número de publicações nessa direção. Para se ter uma idéia do grau de maturidade do processo até então, o passo inicial dessa mineração, ou seja, a identificação das possíveis fontes de informação relevantes dentro de um contexto qualquer, ainda é fruto da criatividade e experiência de especialistas.

A extensa gama de escopos de informação se mostra uma forte barreira para concordância e unificação de paradigmas. A conseqüência direta disso é a falta de soluções automatizadas, desencorajando investimentos pela simples indefinição sobre o consumo de tempo e dinheiro necessários à pesquisa até que esta obtenha resultados justificáveis.

1.1 O PROBLEMA DE PESQUISA

Um dos principais problemas nas redes de distribuição de energia elétrica é o fato dessa área envolver fatores fora de total administração por parte das empresas concessionárias do serviço de distribuição, como, por exemplo, imprevisibilidade de mudanças metereológicas, variações repentinas na demanda de potência, falhas de equipamentos, uso impróprio de

energia (como por exemplo, consumidores declarados como "residenciais" que utilizam equipamentos comerciais ou industriais), uso ilegal de eletricidade (ligações clandestinas para furto de energia), etc. Mas, entre os aspectos que podem ser previstos e posteriormente gerenciados e otimizados, encontram-se as perdas de energia, as avarias em equipamentos causadas por sobrecarga de potência ou tensão, a otimização dos processos de manutenção, a ociosidade de kVAs, entre outros.

No contexto do setor de energia elétrica no Brasil já existem iniciativas para a criação de *data warehouse*s – como é o caso da COPEL (Companhia Paranaense de Energia) em 2004 –, porém, quanto às tecnologias de suporte às estratégias operacionais, esta pesquisa é feita em uma área ainda relativamente nova (TODESCO et al., 2004a). Embora uma grande quantidade de dados esteja sendo armazenada já há muito tempo, o grau de refinamento dessas informações permanece no nível dos sistemas de software operacionais, às vezes dispersos pelos setores da organização e, em outros casos, sem relevante utilidade estratégica.

Especificamente no setor de distribuição elétrica, um dos maiores fatores de motivação para a melhora dos serviços é representado pelos indicadores de confiabilidade e continuidade, definidos e supervisionados pela Agência Nacional de Energia Elétrica (ANEEL).

1.2 OBJETIVOS DO TRABALHO

Este trabalho descreve a pesquisa e a aplicação de técnicas de *Data Mining* em conjunto com Algoritmos Genéticos (AGs) em uma rede de baixa tensão, objetivando encontrar regras de classificação que contribuam para o processo estratégico e de tomada de decisão em empresas distribuidoras de energia elétrica. Tendo obtido resultados significativos, pretende-se por fim disponibilizar aos usuários especialistas as ferramentas e métodos utlizado para a extração das regras.

Durante o trabalho, como parte da revisão bibliográfica, serão vistos:

- os aspectos gerais de sistemas elétricos e da empresa-alvo deste estudo, as Centrais Elétricas de Santa Catarina (CELESC);

- os conceitos de *data warehouse*, fábrica de informações e *Data Mining*;
- as técnicas de mineração de dados com enfoque sobre o reconhecimento de padrões;
- as tarefas de mineração de dados com ênfase sobre a obtenção de regras;
- os Algoritmos Genéticos;
- a análise da adequação de Algoritmos Genéticos no contexto de *Data Mining*; e
- a análise comparativa entre técnicas de IA para otimização e busca.

1.3 METODOLOGIA

O processo de busca das regras coletará amostras, preparará os dados, configurará as possíveis ferramentas utilizadas e realizará experimentos em interação com especialistas no escopo do problema. As atividades serão acompanhadas e, sempre que possível, validadas por eles.

Destaca-se que, embora este trabalho seja abordado do ponto de vista científico, através da aplicação de tecnologias e técnicas pesquisadas, ele não deixa de enfocar o aspecto empresarial do problema, em que qualquer solução de software exige desempenho, confiabilidade, qualidade e robustez.

Portanto, o algoritmo genético selecionado para gerar as regras de classificação foi testado quanto à performance, autonomia e facilidade de interação. Ele também foi comparado com outras soluções de software para serem analisados seus benefícios e seus possíveis pontos fracos.

1.4 JUSTIFICATIVA

Propõe-se encontrar padrões de comportamento na rede elétrica de baixa tensão que permitam prever problemas e falhas técnicas, auxiliando a manutenção da rede e o projeto de circuitos. Quanto à contribuição deste trabalho, apresentam-se dois aspectos fundamentais:

- o valor teórico para a Ciência da Computação (área de Sistemas de Informação), na medida em que revisa técnicas de *Data Mining* para a busca de soluções complexas, aplica de forma cooperativa a Mineração de Dados aliada à Computação Evolucionária e implementa rotinas que fazem uso de Algoritmos Genéticos; e
- o valor prático para o setor de distribuição de energia elétrica, quanto a melhorar a capacidade de previsão de problemas nos circuitos elétricos de baixa tensão, diminuindo assim os custos corretivos em favor da manutenção preventiva, além de criar simples padrões que possam servir como base para o projeto da rede, e abrindo caminho para novas pesquisas de TI nessa área de aplicação.

1.5 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em sete capítulos. Nos três capítulos a seguir é feita a revisão bibliográfica, que serve de base para a proposta aqui levantada: inicialmente as redes de distribuição são descritas em suas principais características; em seguida, aborda-se o *data warehouse* e as técnicas de mineração de dados; por último, apresentam-se de modo geral as definições e os aspectos de algoritmos genéticos. No Capítulo cinco aplica-se um estudo de caso utilizando um algoritmo genético para a geração de regras de classificação. Os resultados do experimento com o AG são apresentados no Capítulo seis, demonstrando o potencial da mineração de dados na extração de conhecimento em redes elétricas de baixa tensão, discutindo as abordagens colocadas em prática e comparando, no contexto do problema, as diversas características das soluções encontradas. Por fim, o último capítulo apresenta as conclusões do estudo e os futuros trabalhos.

2 REDES DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Neste capítulo, o domínio do ambiente da aplicação é apresentado do ponto de vista elétrico e quanto aos aspectos físicos envolvidos na distribuição de energia elétrica. Pretende-se assim descrever brevemente as principais características dos sistemas elétricos, bem como o seu comportamento e as regulamentações impostas pelas agências competentes – sempre com ênfase nas redes de distribuição de energia elétrica.

2.1 CONCEITOS FUNDAMENTAIS DOS SISTEMAS ELÉTRICOS

Esta seção introduz de modo bastante simples alguns dos conceitos principais sobre sistemas elétricos, como, por exemplo, grandezas físicas, instrumentos de medição, materiais, equipamentos utilizados pelas concessionárias[2] de energia elétrica. O objetivo é facilitar o entendimento a respeito dos principais atributos com os quais o algoritmo de mineração de dados que será utilizado deverá trabalhar.

2.1.1 Energia

Segundo Creder (1991), a energia é a potência dissipada ao longo do tempo. Também temos a seguinte definição dada pelo autor: “Tudo aquilo que é capaz de produzir calor, trabalho mecânico, luz, radiação, etc.”.

A energia elétrica é um tipo especial de energia, utilizada para transmitir e transformar a energia primária da fonte produtora, que aciona os geradores nos tipos de energia consumidos em residências. Eletricidade ainda pode ser definida como uma energia intermediária entre a fonte produtora e a aplicação final.

[2] Concessionária ou permissionária é definida como o agente titular de concessão ou permissão federal para explorar a prestação de serviços públicos de energia elétrica (ANEEL, 2001).

2.1.2 Corrente elétrica

É também chamada de Amperagem, pois a sua unidade de medida é o Ampère (Amp). De acordo com Creder (1991), corrente elétrica é:

> O deslocamento de cargas dentro de um condutor, quando existe uma diferença de potencial elétrico entre as suas extremidades. Tal deslocamento procura restabelecer o equilíbrio desfeito pela ação de um campo elétrico ou outros meios (reação química, atrito, luz, etc.).

Existem dois tipos básicos de corrente (CREDER, 1991):

1) Corrente contínua: não varia ao longo do tempo; e
2) Corrente alternada: oscilatória, que varia de amplitude em relação ao tempo segundo uma lei definida.

Em relação à corrente têm-se outros conceitos importantes: a) *freqüência* é definida como o número de vezes por segundo em que a corrente alternada completa um ciclo (MUSEUM OF SCIENCE BOSTON, 2005). A unidade de medida da freqüência é o Hertz (ciclos por segundo); b) *resistência* é a medida da dificuldade encontrada pela corrente de passar através de um dado elemento (MUSEUM OF SCIENCE BOSTON, 2005). A unidade de medida da freqüência é o Ohm (Ω).

2.1.3 Tensão

Tensão é a diferença de potencial entre dois pontos de um campo eletrostático (1991). Nas redes de distribuição elétrica a tensão é classificada para efeito de consideração de acordo com vários aspectos. Embora todas as definições de tensão sejam expressas em volts (V) ou quilovolts (kV), de acordo com a ANEEL (2000), distinguem-se:

- Tensão Nominal (TN): valor eficaz de tensão pelo qual o sistema é projetado;

- Tensão de Atendimento (TA): valor eficaz de tensão no ponto de entrega ou de conexão, obtido por meio de medição, podendo ser classificada em adequada, precária ou crítica, de acordo com a leitura efetuada;

- Tensão Contratada (TC): valor eficaz de tensão que deverá ser informado ao consumidor por escrito ou estabelecido em contrato;

- Tensão de Leitura (TL): valor eficaz de tensão, integralizado a cada 10 (dez) minutos, obtido de medição por meio de equipamentos apropriados;

- Tensão Não Padronizada (TNP): valor de tensão nominal;

- Tensão Nominal de Operação (TNO): valor eficaz de tensão para o qual o sistema é designado.

2.1.4 Potência

Trata-se da energia aplicada por segundo para realizar atividades. Mede-se a potência em Watts. Na área elétrica, potência é o produto da tensão pela corrente. Em circuitos de corrente alternada, existem três tipos de potência (CREDER, 1991):

1) Potência ativa: é a potência dissipada em calor;
2) Potência reativa: potência trocada entre gerador e carga sem ser consumida;
3) Potência aparente: soma vetorial das duas potências anteriores.

2.1.5 Instrumentos de medição

O instrumento mais comum de medição de energia elétrica é o registrador. Ele funciona através dos campos de corrente elétrica gerados por bobinas de corrente e de potencial induzindo a rotação de um disco, o qual está ligado a um registrador.

2.1.6 Equipamentos

A seguir são descritos alguns dos principais equipamentos que compõem as redes de distribuição elétrica.

a) Alimentador: todo circuito primário ligado diretamente ao circuito secundário de uma subestação de distribuição, possibilitando a alimentação direta dos transformadores e pontos de consumo sob a mesma tensão do referido circuito (CELESC, 1980).

b) Condutor: material com baixa resistência elétrica que permite à eletricidade se mover facilmente através dele (MUSEUM OF SCIENCE BOSTON, 2005).

c) Subestação: parte das instalações elétricas da unidade consumidora atendida em tensão primária que agrupa os equipamentos, condutores e acessórios destinados à proteção, medição, manobra e transformação de grandezas elétricas (ANEEL, 29 nov. 2000); também roteia e administra o fluxo elétrico (PUBLIC POWER COUNCIL, 2005), modificando o nível de tensão para torná-lo apropriado ao consumidor final (CENTRAL VERMONT PUBLIC SERVICE, 2005).

d) Trecho: o espaço de transmissão elétrica conectado por dois pontos elétricos.

e) Transformador de tensão: dispositivo que transforma a tensão elétrica de um nível para outro; pela potência de saída não poder exceder a potência de entrada, a corrente final é reduzida em proporção direta ao ganho de voltagem (MUSEUM OF SCIENCE BOSTON, 2005).

2.2 ELEMENTOS BÁSICOS DOS SISTEMAS ELÉTRICOS

Para compreender como funciona a baixa tensão elétrica, é preciso saber que esta se situa dentro de uma estrutura elétrica mais complexa. Assim, Creder (1991) divide o processo necessário para o funcionamento de um sistema elétrico através de três componentes:

1) produção;
2) transmissão; e
3) distribuição.

A seguir cada um desses componentes é brevemente definido e explicado.

2.2.1 Produção

A geração de energia elétrica é obtida através do uso da energia potencial da água (hidroelétrica), de combustíveis (termoelétrica) ou mecânica (cinética). Os combustíveis podem ser fósseis (petróleo, carvão, etc.), não fósseis (madeira, por exemplo) ou nuclear (urânio enriquecido) (CREDER, 1991).

É interessante saber que as companhias concessionárias de energia não necessariamente produzem a eletricidade que distribuem, tampouco precisam ser consumidoras de somente uma companhia geradora.

2.2.2 Transmissão

A transmissão é a estrutura responsável por conectar as companhias produtoras de energia e as companhias distribuidoras de eletricidade. Quanto a essa parte do sistema elétrico, ainda segundo Creder (1991): “Transmissão significa o transporte da energia elétrica gerada até os centros consumidores”.

2.2.3 Distribuição

De acordo com Creder (1991), a distribuição da energia elétrica é a parte que ocorre já dentro dos centros urbanos, começando na subestação abaixadora – onde a tensão da linha é transformada em valores padrões da rede primária (alta e média tensão) – e seguindo até a

subestação abaixadora para a baixa tensão – modificando a tensão para alimentar a rede secundária, ou seja, ao nível de utilização (baixa tensão).

É sobre os dados e as informações gerados durante a fase de distribuição que esse trabalho se aplica, mais especificamente sobre a distribuição da baixa tensão.

2.3 CLASSIFICAÇÃO DOS CONSUMIDORES

Inicialmente é preciso fazer a distinção dentro da terminologia adotada pela ANEEL para referenciar o consumidor e a unidade consumidora de energia elétrica. De acordo com a ANEEL (2000), um consumidor é definido como sendo:

> Pessoa física ou jurídica, ou comunhão de fato ou de direito, legalmente representada, que assumir a responsabilidade pelo pagamento das faturas de energia elétrica e pelas demais obrigações fixadas em normas e regulamentos da ANEEL, assim vinculando-se ao contrato de fornecimento, de uso e de conexão ou de adesão, conforme cada caso. (ANEEL, 2000, Art. 2º, § 3º).

Já a Unidade Consumidora (UC) é conceituada como a representação de:

> Um conjunto de instalações de equipamentos elétricos caracterizado pelo recebimento de energia elétrica em um só ponto de entrega, com medição individualizada [...]. (ANEEL, 2000, Art. 2º, § 40º).

A mesma resolução estabelece a conformidade dos níveis de tensão de energia elétrica em regime permanente[3] para os tipos de consumidor atendidos conforme a tensão nominal a eles distribuída.

1) UC de alta-tensão: maior ou igual a 69 kV;
2) UC de média tensão: maior que 1 kV e menor que 69 kV;
3) UC de baixa tensão: igual ou inferior a 1 kV.

[3] Trata-se do intervalo de tempo da leitura de tensão, em que não ocorrem distúrbios elétricos capazes de invalidar a leitura, definido como sendo de dez minutos (ANEEL, 2000).

A tensão é distribuída em até três fases (corrente) paralelas às UCs de baixa tensão. Assim, existem consumidores monofásicos, bifásicos e trifásicos.

Existem ainda outros tipos de classificação que podem ser atribuídos aos consumidores, conforme as atividades que exercem ou ainda as condições em que se encontram. A seguir são citadas as principais classes de consumidor consideradas na distribuição de energia.

- Residenciais.
- Comerciais.
- Industriais.
- Rurais.
- Públicos.
- Serviços essenciais[4].

2.4 QUALIDADE NA DISTRIBUIÇÃO

A qualidade do atendimento aos consumidores de energia elétrica no Brasil é padronizada e fiscalizada pela agência ANEEL (Agência Nacional de Energia Elétrica), conforme resoluções e decretos federais. Devido às penalidades previstas em lei para controlar a qualidade da eletricidade recebida nas UCs, as concessionárias de energia elétrica (EE) precisam gerenciar a rede de distribuição, procurando, de modo geral, evitar ou pelo menos diminuir a freqüência e a duração das interrupções no fornecimento de energia.

Nesta seção é feita uma breve revisão sobre as resoluções da ANEEL que descrevem o cálculo dos indicadores de continuidade e confiabilidade e que padronizam os níveis de tensão para os diversos tipos de consumidor.

[4] Serviço ou atividade considerada como de fundamental importância para a sociedade, por exemplo, hospitais, companhias de tratamento de água e esgoto, lixo, telecomunicações, tráfego aéreo, segurança pública, etc. (ANEEL, 2001).

2.4.1 Indicadores de continuidade

Os indicadores representam de forma quantitativa a qualidade na distribuição de energia elétrica da concessionária. Através deles, a continuidade na distribuição de EE, seja coletiva ou individualmente às unidades consumidoras, é supervisionada e avaliada comparativamente ao chamado "padrão de continuidade" – valor máximo definido para um indicador (ANEEL, 2001). Trata-se de uma maneira simples de mensurar os problemas ocorridos no atendimento elétrico. Para entender o conceito de continuidade, é preciso compreender a definição de interrupção. Na prática existem quatro tipos de interrupção, segundo os quais uma "descontinuidade" pode ser classificada, são eles:

1) *Interrupção*: descontinuidade do neutro ou da tensão disponível em qualquer uma das fases de um circuito elétrico que atende à UC;

2) *Interrupção de longa duração*: toda interrupção do sistema elétrico com duração maior ou igual a um minuto;

3) *Interrupção programada*: interrupção prevista, com um tempo preestabelecido e previamente avisada[5], com o objetivo de intervenção (manutenção, modificação, nova implementação, etc.) no sistema elétrico da concessionária;

4) *Interrupção de urgência*: interrupção deliberada no sistema elétrico, que, devido ao aspecto de urgência na execução de serviços, não oferece possibilidade de aviso prévio ou de agendamento.

Segundo a ANEEL (2001), os indicadores devem ser calculados pelas companhias elétricas. A fórmula utilizada para se encontrar cada indicador pode ser encontrada na própria resolução da Agência. Até janeiro de 2005, para o cálculo de indicadores em geral as concessionárias estavam obrigadas por contrato a considerar:

1) interrupções com duração maior ou igual a três minutos; ou

[5] Conforme o tipo de consumidor (industrial, comercial, que presta serviço ou requer serviços essenciais, etc.), o aviso sobre a data, o horário de início e fim deve ser dado por meios diferentes e com antecedência variada (ANEEL, 2001).

2) interrupções com duração maior ou igual a um minuto.

No entanto, a partir da data citada, qualquer interrupção no fornecimento de EE superior a um minuto já passa a valer para o cálculo dos indicadores. Mesmo sendo tão exigente, há exceções razoáveis, como, por exemplo, quando obras de interesse exclusivo do consumidor causam *black-out* na UC; nesse caso, a concessionária não computa o problema para os indicadores. A Tabela 1.1 apresenta os indicadores definidos pela ANEEL.

DEC - Duração Equivalente de Interrupção por Unidade Consumidora	Média de intervalo de tempo em que cada UC do conjunto considerado, no período de observação, sofreu descontinuidade da distribuição de energia elétrica.
FEC - Freqüência Equivalente de Interrupção por Unidade Consumidora	Média do número de interrupções ocorridas, no período de observação, em cada UC do conjunto considerado.
DIC - Duração de Interrupção Individual por Unidade Consumidora	Intervalo de tempo em que cada UC, no período de observação, sofreu descontinuidade da distribuição de energia elétrica.
FIC - Freqüência de Interrupção Individual por Unidade Consumidora	Número de interrupções ocorridas, no período de observação, em cada UC.
DMIC - Duração Máxima de Interrupção Contínua por Unidade Consumidora	Tempo máximo de interrupção contínua da distribuição de energia elétrica em uma UC qualquer.
Indicador de Continuidade	Representação numérica do desempenho de um sistema elétrico. É utilizado para a mensuração da continuidade alcançada e análise comparativa com os padrões estabelecidos.
Indicador de Continuidade Global	Representação numérica do desempenho de um sistema elétrico agrupado por empresa, estado, região ou país.

Tabela 1.1 - Indicadores de Confiabilidade ANEEL

FONTE: ANEEL, 2001.

2.4.2 Metas de continuidade

São os valores máximos estabelecidos para cada indicador de continuidade. As metas são mensais, trimestrais e anuais nos períodos correspondentes ao ciclo de revisão das tarifas, conforme resolução específica (ANEEL, 2001).

2.4.3 Avaliação da tensão

A tensão é analisada de diversas formas na rede de distribuição, desde a regulagem dos equipamentos, a perda devida ao tipo e à extensão do condutor, até sua participação na demanda de potência. Para se avaliar a tensão na rede de distribuição, há dois procedimentos padrões regulamentados pela ANEEL (2000).

O primeiro procedimento se dá por meio da avaliação trimestral exigida pela ANEEL. Nesse caso, uma amostra de UCs escolhidas por critério aleatório estatístico é entregue à companhia concessionária de energia com sessenta dias de antecedência até a entrega dos resultados. A companhia faz as medições utilizando equipamentos conforme especificações e transforma os valores encontrados nos seguintes indicadores (identificados por UC):

a) DRP: Duração Relativa da Transgressão de Tensão Precária;
b) DRC: Duração Relativa da Transgressão de Tensão Crítica.

O segundo procedimento para medição dos níveis de tensão diz respeito à reclamação de consumidores, o qual se chama Pedido de Verificação do Nível de Tensão (PVNT). Se houver reclamações quanto ao nível de tensão de atendimento (conforme descrito na seção 2.1.3) por parte do consumidor, a ANEEL (2000) regulamenta os procedimentos que devem ser tomados pela companhia de distribuição que atende àquela UC.

2.4.4 Penalidades

A penalidade para a violação da continuidade do fornecimento de EE é uma compensação concedida ao consumidor cujo valor é creditado na fatura de energia elétrica no mês subseqüente à apuração dos indicadores da ANEEL (2001).

Para o cálculo da penalidade quanto à continuidade individual (de cada UC), a média aritmética (CM) dos valores líquidos das faturas ou dos encargos correspondentes aos meses em que houve violação é multiplicada por:

- 10 em caso de violação de padrão mensal;
- 30 quando o padrão trimestral for ultrapassado;
- 120 para a quebra no padrão anual.

As multas individuais são descontadas das multas daquele conjunto, de acordo com um cálculo de compensação fornecido em resoluções da ANEEL (2001). Para as violações do conjunto, a multa é paga à ANEEL conforme a Resolução n.º 318 (ANEEL, 1998).

2.5 CONTEXTO DA APLICAÇÃO NA REDE DE DISTRIBUIÇÃO ELÉTRICA

2.5.1 Contexto de aplicação

Este trabalho foi desenvolvido baseando-se no ambiente de informação da companhia CELESC (Centrais Elétricas de Santa Catarina), uma concessionária do serviço público brasileiro de energia elétrica, fundada em 1955. Abrangendo 262 municípios em SC e um no Paraná (CELESC, 2003) (além de 25 outros municípios indiretamente), a CELESC atende hoje a 1.927.916 clientes e é responsável pela distribuição de energia em 91,79% da área do Estado de Santa Catarina (CELESC, 2004).

2.5.2 Ambiente de dados corporativo

A CELESC divide a rede de distribuição elétrica dentro do Estado em 16 grandes áreas. Em cada uma delas, a Companhia possui uma agência regional, a qual é responsável por coletar, armazenar e replicar informações, além, é claro, de prestar serviços aos consumidores da respectiva área.

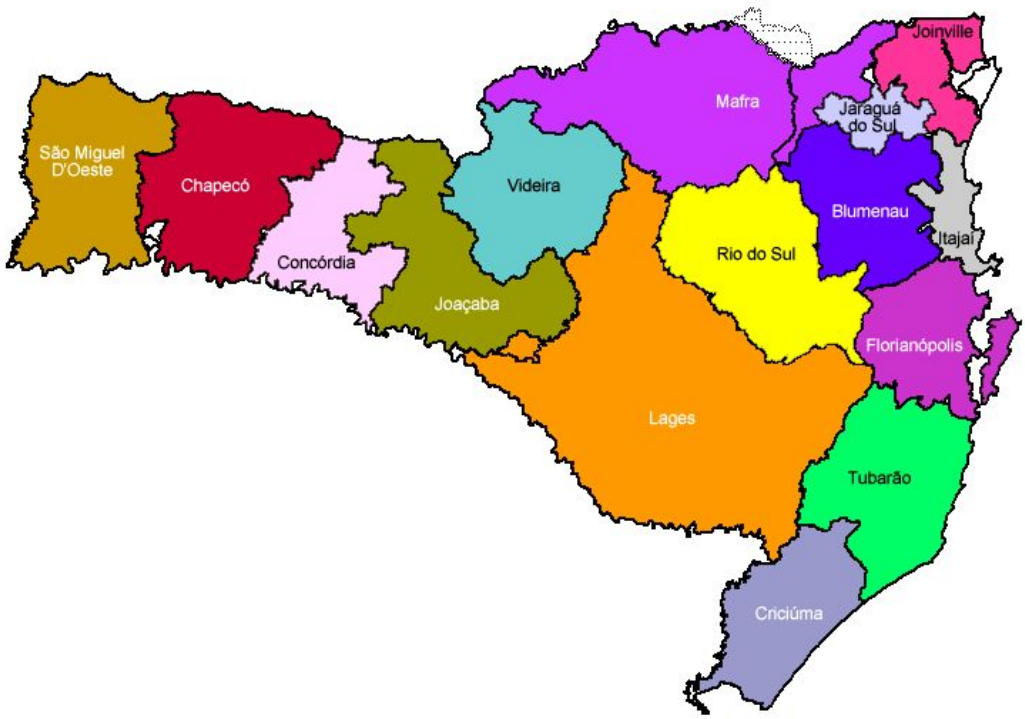

Figura 2.1 - Divisão política das agências regionais da CELESC

Cada região é comumente chamada pelo nome da cidade onde se encontra o escritório – cidade de maior influência naquela área –, por exemplo, “Agência Regional de Joinville”. A divisão política das agências regionais está descrita na Figura 2.1.

Os sistemas corporativos da empresa reúnem dados históricos, sumarizações e dados detalhados sobre toda a rede de distribuição de energia no Estado. Tais sistemas estão distribuídos pela maioria das regionais onde contêm dados relativos àquela regional em que se encontram.

As agências que não possuem bancos de dados em suas respectivas cidades-sede armazenam seus dados no escritório central de administração da companhia, na capital do Estado, Florianópolis. Além desses dados, a administração também centraliza muitas das informações da companhia.

O ambiente computacional para apoio aos processos gerenciais dentro da empresa é formado basicamente por dois principais sistemas de informação: SIMO (Sistema Integrado de Manutenção e Operação) e GENESIS (Gerência Integrada de Sistemas de Distribuição de Energia Elétrica).

O sistema SIMO, entre outras funções, registra os problemas ocorridos na rede de distribuição, os dados sobre a manutenção desses problemas, as informações sobre a interrupção no fornecimento de energia, as reclamações de consumidores, etc. Este sistema atua dando suporte de informações principalmente aos atendentes de plantão na companhia, às equipes de manobra, planejamento e execução de manutenção na rede elétrica (TODESCO et al., 2004c).

O sistema GENESIS armazena informações cartográficas (organização urbana em torno da rede – ruas, edificações, disposição do circuito, etc.), grandezas elétricas (como, por exemplo, carregamento de potência, queda de tensão, tensão em cada fase elétrica), estrutura física (equipamentos, postes, cabos) e características topológicas da rede de distribuição, tanto primária quanto secundária (TODESCO et al., 2004a).

2.5.3 Metodologia para a obtenção de dados sobre equipamentos

Para estimar e validar as regras de classificação encontradas quanto à sua capacidade de predição, exige-se que a base de dados contenha um período histórico de informações, o qual poderá prover o suporte e a confiança da regra (item 1.1.3.3.10) a partir de análise através de períodos de tempo. Porém, para que as regras, uma vez postas em operação, sejam capazes de “padronizar” problemas é necessário também analisar as fontes dos dados em relação à sua validade no tempo, à rigidez dos cálculos que alimentam essas fontes, à confiabilidade dos métodos utilizados, bem como à regularidade com que elas são atualizadas.

A escolha entre as possíveis abordagens para efetuar a medição das grandezas elétricas baseia-se principalmente na viabilidade econômica das alternativas. Atualmente, a coleta dos dados não é feita apenas por meio da medição em campo dos equipamentos, mas também são feitos cálculos elétricos que estimam a potência demandada e a tensão (entre outros aspectos) desses equipamentos da rede. Tais cálculos são realizados de forma indireta, baseando-se no consumo apontado na fatura elétrica mensal de cada consumidor ligado àquele circuito da rede.

Esse método gera uma aproximação dos reais valores, mas sem dúvida não é a forma mais precisa para obter os dados (TODESCO et al., 2004c). No entanto, a confiabilidade desses dados é válida, isto é, eles possuem margem de erro aceitável, já que são utilizados pela companhia em outras atividades e nos sistemas de informação corporativos.

Existe ainda outro método que não exige que a medição seja feita diretamente por técnicos eletricistas. Porém, tal metodologia requer a modificação dos atuais medidores de energia dos consumidores, bem como a instalação de um sistema de teleleitura nos transformadores, o qual faria varreduras no equipamento e transmitiria os dados elétricos coletados através de ondas portadoras (por exemplo, VHF) (TODESCO et al., 2004c). O método de teleleitura já é bastante utilizado em equipamentos de média e alta-tensão em subestações transformadoras.

2.5.4 Possibilidades de aplicação de business intelligence na área de distribuição de energia elétrica

Este estudo inseriu-se no contexto de um projeto de P&D cujo objetivo principal é conceber e implantar uma plataforma de gestão que organiza, em um único ambiente, informações relativas ao projeto de redes de distribuição de energia, manutenção, realização de obras e operação do sistema, e comercialização de energia. Para isso, aplicam-se técnicas de Data Warehousing e de Mineração de Dados. Tal plataforma de gestão visa subsidiar análises de cenário, prover acompanhamento contínuo da qualidade do fornecimento, reduzindo custos, otimizando processos operativos e de tomada de decisão e, conseqüentemente, melhorando a qualidade do atendimento aos clientes quanto ao fornecimento de energia (TODESCO et al., 2004a).

Diante das exigências impostas pela Agência nacional que regula o setor elétrico no Brasil (ANEEL) em relação à qualidade na distribuição de energia, as empresas de energia elétrica no País têm buscado constantemente melhorar os serviços prestados (TODESCO et al., 2004a). Além disso, Todesco et al. (2004a) acrescentam que a preocupação com a evolução dos serviços tem, como fator motivador, a crescente competitividade e, como fator condicionante, a sobrevivência da empresa no mercado.

O ambiente estruturado pelo projeto citado para mineração de dados fornece diversas perspectivas quanto à descoberta de conhecimento aplicável, principalmente, considerando-se o fato de que se trata de uma base de dados ainda em desenvolvimento, isto é, pouco explorada por aplicações de mineração de dados. Hoje já existem projetos de pesquisa dentro da CELESC utilizando-se da diversidade, padronização e sumarização dos dados armazenados nos *data marts* até agora desenvolvidos, entre os quais, podem ser citados um sistema especialista (TODESCO et al., 2004b) e um sistema de previsão de demanda de energia (TODESCO et al., 2004c).

A integração entre esses sistemas de informação é desejável visto que a intersecção do conhecimento já alcançado por eles, a realimentação de forma colaborativa e até mesmo a validação inteligente e contínua entre tais sistemas contribuirão inevitavelmente para a sua própria evolução.

3 DATA WAREHOUSE E DATA MINING

Este capítulo introduz os principais conceitos sobre *Data Warehouse* e *Data Mining*, explorando o assunto do ponto de vista das organizações consumidoras de informação e do analista de sistemas. Também é dentro da visão de ambos que a Fábrica de Informações Corporativas é descrita em seus diversos componentes, os quais irão integrar operacionalmente as necessidades organizacionais da corporação à estrutura modelada pelo minerador de conhecimento.

3.1 *DATA WAREHOUSE* (DW)

A importância da informação no apoio à tomada de decisões é indiscutível, principalmente para as grandes organizações. Mas a informação em sua forma bruta não possui real e efetiva utilidade para o processo de gerência e administração. Para isso, é necessário organizá-la, bem como atualizá-la, tratá-la e mantê-la.

A evolução da informática tornou possível que os consumidores de informação – empresas, grandes instituições (governamentais ou não), centros de pesquisa, etc. – tivessem a sua demanda por sistemas de informação atendida quanto aos fatores de armazenamento e apresentação (FAYYAD, 1997). Porém, é surpreendente que muitos dos usuários desses repositórios de dados ainda não tenham se apercebido do potencial que tais repositórios possuem para gerar inteligência e, conseqüentemente, produtividade, lucratividade, avanços, economia, entre outros grandes benefícios.

No entanto, os sistemas transacionais[6] mais comumente utilizados não conseguem assegurar adequadamente consistência, integração e precisão dos dados. Fez-se então necessária a criação de um ambiente de apoio à decisão robusto, sustentável e confiável.

[6] Modelos de bancos de dados projetados para suportar freqüentes transações (operações internas ao BD, conhecidas como *transactions*) de registros numa taxa relativamente alta.

A Figura 3.2 apresenta a estrutura de informação dentro da organização através do uso de um *data warehouse*.

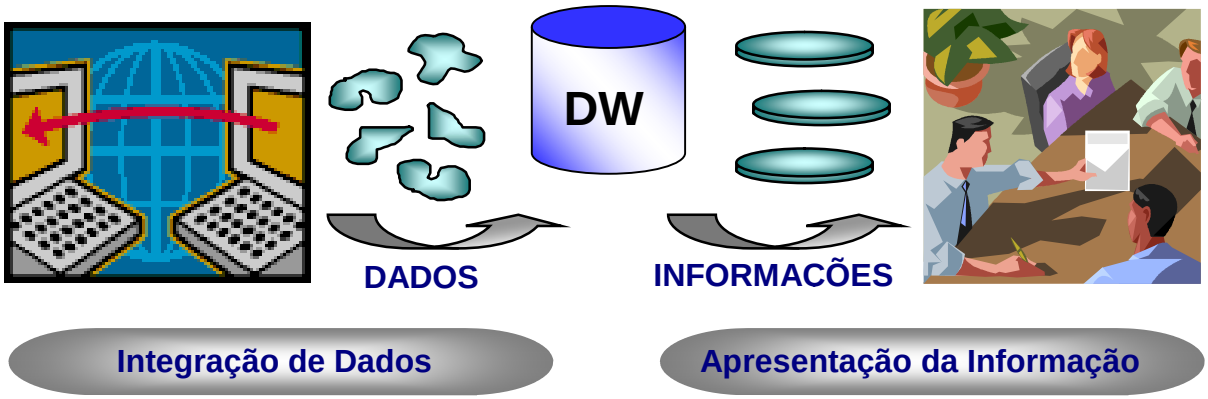

Figura 3.2 - Estrutura de Informação por um *data warehouse*

Observando a dificuldade organizacional, as deficiências do modelo relacional comumente utilizado, a falta de integridade e até a indisponibilidade de acesso à enorme massa de informações existente, o conceito de *data warehouse* surgiu como o primeiro passo na transformação de sistemas de banco de dados. Assim, o DW deixou de ser somente um armazenador confiável para tornar-se uma poderosa ferramenta cuja principal finalidade é o suporte à decisão (FAYYAD, 1997).

Quanto à maneira de modelar um *data warehouse* em comparação a qualquer outro banco de dados, a mais clara definição é dada por Kimball (2002), que estabelece DW como um conglomerado de áreas de apresentação e de estágio (*data staging*) de uma organização, em que o dado operacional é especificamente estruturado para prover performance e facilidade de uso em operações de consultas e análise.

Dizemos resumidamente que *data warehouse* é um conjunto de dados atuais e históricos, extraídos de vários sistemas operacionais, destinados a fornecer informações que auxiliem o processo de tomada de decisão. Assim, um *data warehouse* consiste em organizar os dados corporativos da melhor maneira para fornecer informações aos gerentes e diretores das organizações na tomada de decisão. Tudo isso é feito em um banco de dados paralelo aos sistemas operacionais da empresa.

A tecnologia de um DW difere dos padrões operacionais de sistemas de banco de dados em três principais aspectos:

1) dispõe de habilidade para extrair, tratar e agregar dados de múltiplos sistemas operacionais em *data marts* separados;
2) armazena dados freqüentemente em formato de cubo (OLAP – Online Analytical Process) multidimensional, permitindo rapidamente agregar dados e detalhar análises (*drilldown*); e
3) disponibiliza visualizações informativas, pesquisando, reportando e modelando capacidades que vão além dos padrões de sistemas operacionais freqüentemente oferecidos.

As principais características de um DW podem ser resumidas na definição de Inmon (1997), em que *data warehouse* é um conjunto de dados orientado por assuntos, não volátil, variável com o tempo e integrado, criado para dar suporte à decisão.

Orientado por assuntos, significa que o banco de dados abordará um determinado aspecto dentro da organização real a uma área de negócio (marketing, departamento pessoal, setor comercial, etc.) e sobre a qual será mantida a informação. Cada um desses assuntos pode representar um *data mart* diferente pertencente ao DW.

Segundo Kimball (2002), *data marts* são conjuntos flexíveis de dados, idealmente baseados na maior granularidade possível de se extrair de uma fonte operacional, e apresentados em um modelo simétrico (dimensional) na execução de consultas inesperadas. De forma mais simplificada, pode-se definir *data marts* como a representação de dados de um único processo de negócio, isto é, dados baseados em assuntos específicos. Esses assuntos geralmente representam as diferentes áreas dentro da organização.

A volatilidade refere-se ao fato de o warehouse não sofrer atualizações da maneira convencional, como os demais sistemas tradicionais. Sendo o *data warehouse* um sistema de apoio à decisão, atualizações freqüentes sobrecarregariam a sua capacidade de gerar consultas, pois as suas entidades estariam constantemente sendo alocadas para inserções, alterações e exclusões de registros. Mesmo a sua estrutura, como banco de dados relacional, possui diferenças, diminuindo generalizações/especializações para aumentar o desempenho

em consultas SQL que façam junção entre as tabelas. Basicamente, pode-se dizer que um *data warehouse* tem apenas duas operações: (1) a carga de dados; e (2) a consulta.

A característica do *data warehouse* quanto à perspectiva temporal objetiva tornar possível reproduzir situações da organização em momentos diferentes pelos quais ela passou, armazenando dados históricos para retratar os assuntos ao longo do tempo. Por exemplo, uma empresa gostaria de analisar como se comporta determinado cliente após sua mudança de estado civil, porém alterando o cadastro dele, todas as suas compras não vão poder ser distinguidas nesse aspecto; em um *data warehouse* isso não ocorre, pois os dados sobre o cliente antes e depois da alteração de estado civil são armazenados separadamente; e desse modo se pode escolher entre analisar o mesmo cliente e as compras dele feitas na empresa observando se houve diferença em seu comportamento por comparação.

A integração é a parte mais importante desse processo, pois ela será responsável por unir os dados de vários sistemas existentes na empresa e colocá-los no mesmo padrão. Um DW extrai dados de diversos sistemas da organização (até mesmo de SGBDs diferentes) ou dados externos. O processo de popular um DW é conhecido como ETL (Extração, Transformação e Carga), em que os dados são:

1) extraídos de bancos de dados, de arquivos, da Internet, etc. para uma área de estágio (área temporária);
2) formatados e convertidos em um único padrão; e
3) carregados no *data warehouse*.

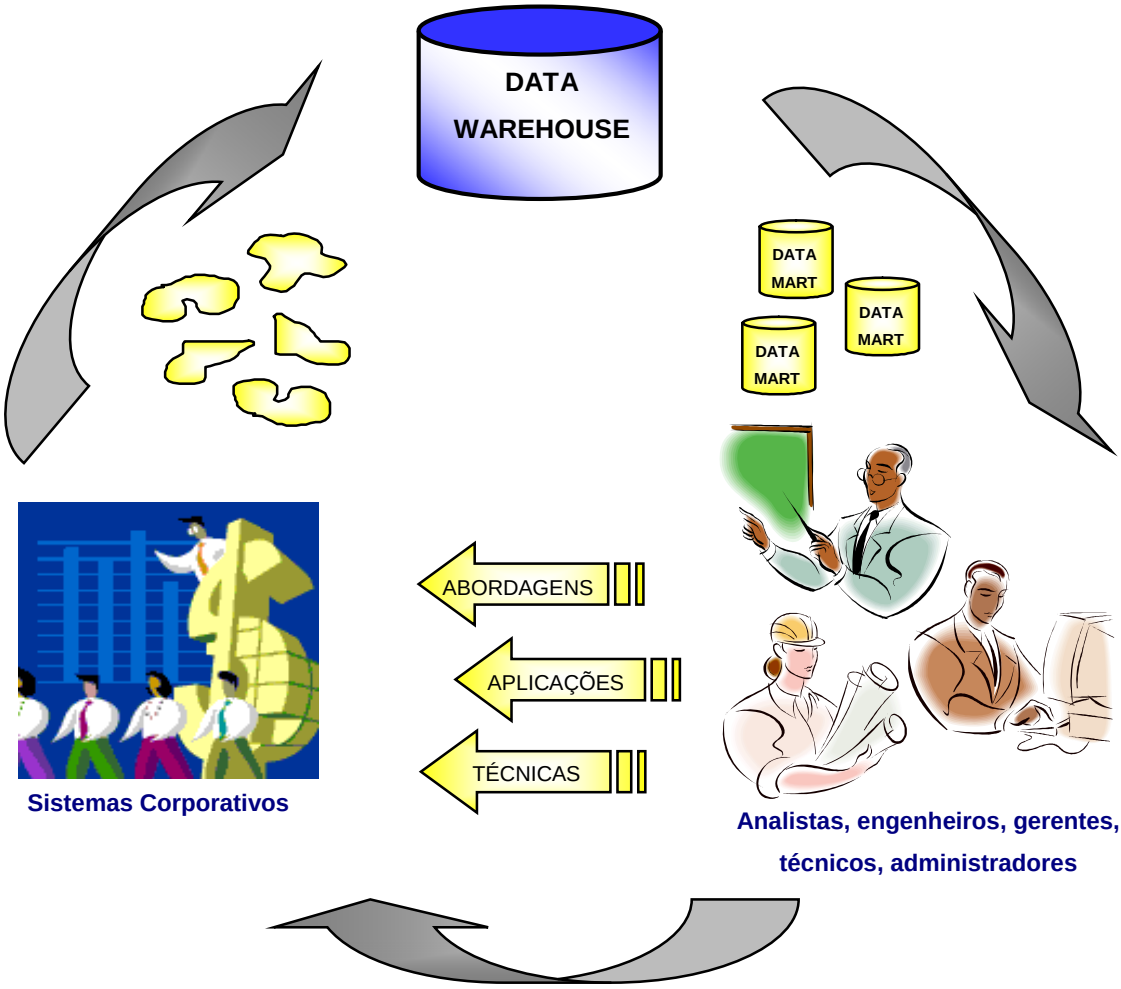

Figura 3.3 - Fluxo de conhecimento utilizando *data warehouse*

A Figura 3.3 apresenta, de maneira geral, o fluxo de conhecimento dentro de uma organização utilizando *data warehouse* em seu processo estratégico. O fluxo começa pela extração e integração dos dados a partir de bancos de dados dos sistemas de software operacionais (controle financeiro, gerência de produção, administração de recursos, etc.). Em seguida, apresenta aos usuários as informações a partir dos *data marts* que compõem o DW, conforme os diferentes tipos de consumidor de informação da corporação. E, por fim, realimenta o fluxo do conhecimento dentro da organização através da tomada de decisão quanto a novas abordagens estratégicas, à aplicação e ao investimento de recursos, à descoberta de melhores técnicas, entre outros.

3.2 FÁBRICA DE INFORMAÇÕES CORPORATIVAS (CORPORATE INFORMATION FACTORY - CIF)

Com a evolução da tecnologia de informação, diversos novos conceitos dentro da disciplina Banco de Dados surgiram nas últimas décadas, a maioria deles com o intuito de apoiar o processo de apresentação e análise. Assim, a antiga imagem que fazíamos de uma arquitetura de banco de dados para extração de informações hoje está bem distante da realidade.

A complexidade inerente à extração de conhecimento, a qual deve existir paralelamente às atividades transacionais, exigiu uma estrutura não apenas cooperativa, mas que poderíamos comparar a um processo de mutualismo[7], em que em um mesmo ambiente existem entidades que alimentam com dados outras entidades, as quais, por sua vez, geram dados de controle para administração das primeiras. Essa infra-estrutura de informações é proposta por Inmon et al. (2001) e é conhecida como Fábrica de Informações Corporativas. Sua arquitetura e seus componentes são apresentados e descritos a seguir, de acordo com o autor, na Figura 3.4.

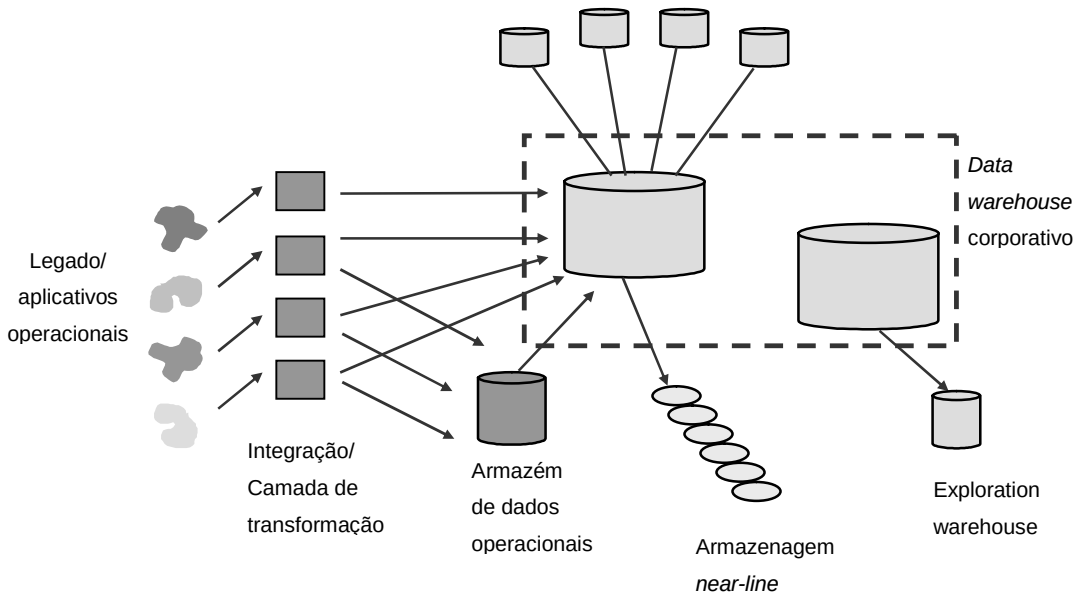

Figura 3.4 - A infra-estrutura por trás da informação: CIF

FONTE: INMON et al., 2001.

[7] Tipo de associação entre organismos de espécies diferentes e na qual há benefícios para uns e outros.

3.2.1. O ambiente de aplicativos de legado/operacionais

Trata-se do ambiente em que os sistemas de negócio coletam dados detalhados dos usuários. Caracteriza-se pela realização de uma atividade principal denominada Transação (inserção, alteração e exclusão), motivo pelo qual também é conhecido como “ambiente transacional”. Nesse contexto normalmente não há integração e nem consenso sobre as entidades de negócios.

3.2.2. A camada de integração e de transformação

Nesta camada os dados coletados a partir de diferentes aplicativos são convertidos e transformados para alcançar sua padronização (de domínio, unidades e tipos de dados). Já existem softwares para executar essa integração entre o ambiente operacional e a camada de integração (GONÇALVES, 2003), efetuando inclusive a documentação dos tipos de transformação e os mapeamentos programados pelo responsável por modelar o DW.

3.2.3. O *data warehouse* corporativo

O *data warehouse* corporativo, ou simplesmente *data warehouse*, armazena os dados limpos, transformados e integrados, conforme descrito anteriormente. Sua estrutura permite dados em formato granular, resumido ou agregado, e sua característica histórica é vital para busca de padrões, funções de regressão ou qualquer outra análise de tendências ao longo do tempo.

3.2.4. Os múltiplos *data marts*

Existem duas abordagens diferentes de projeto: a *top-down* (o DW é dividido em áreas menores) e a *bottom-up* (os *data marts* independentes são construídos aos poucos, e o conjunto deles resultará em um DW). O grande benefício de se construírem *data marts* aos

poucos em vez de todo o *data warehouse* diz respeito ao seu custo e aos prazos, os quais são inferiores ao de um DW.

3.2.5. O *Exploration Warehouse* (EW)

Como propósito geral, o conceito de DW foi desenvolvido para apoiar o processo de tomada de decisão, isto é, o processamento analítico através de ferramentas de *front-end*. De fato, as operações OLAP somente são válidas e grandemente eficazes no ambiente de um DW. Considerando-se essas mesmas características que tornam um *data warehouse* propício a consultas refinadas (detalhadas e históricas), as técnicas de *data mining* também encontraram nele um ambiente favorável à exploração e extração de conhecimento.

No entanto, com as aplicações analíticas se popularizando dentro da organização, mesmo um DW, livre de ações transacionais como as bases operacionais, está sujeito a uma grande carga de processamento, nesse caso um processamento analítico. Embora a freqüência de acessos seja menor em um DW, consultas que buscam agregações ou que "fatiam" os dados levam um tempo consideravelmente grande para serem executadas (precisando até ser programadas[8] no BD em algumas situações específicas, por exemplo, um resumo periódico envolvendo várias entidades de negócio externas).

O problema de desempenho do DW tende a aumentar conforme cresce o número de sistemas analíticos e também de usuários que utilizam esse tipo de sistema. Isso se deve ao fato de que o DW, diferente de *data marts* e bases operacionais, representa uma fonte única para toda a organização, ou seja, independentemente de setor e área dos consumidores de informação, todas as aplicações de análise irão consultar o mesmo local: o *data warehouse*.

A mineração de dados possui demanda de processamento mais alta ainda do que as consultas realizadas pelas ferramentas de *front-end* do DW. Tal aspecto se deve ao seu caráter exploratório, mais comum do que o focalizado, em que a abrangência de dados em níveis granulares e globalmente quanto às entidades é inevitável. Além disso, quando o processo de

[8] Consultas programadas são consultas agendadas no SGBD para serem executadas e entregues em determinado horário ao cliente do BD que as solicitou.

data mining possui um contexto particular já conhecido, não se faz necessário trabalhar em um ambiente que contenha toda uma variedade de dados existentes e possíveis.

Para evitar ou resolver problemas de desempenho no DW causados por sobrecarga, surgiu o conceito de *Exploration Warehouse*, cuja estrutura é desenvolvida com o objetivo específico de prover um ambiente para exploração de informações. Ele deriva do DW corporativo na medida em que é alimentado por este; sua principal diferença de um DW corporativo está no propósito para o qual foi modelado: pesada análise de informação detalhada e histórica ainda inexplorada (INMON et al., 2001).

O EW geralmente é uma estrutura temporária, visto que não há atualizações constantes previstas para ele (semelhante às cargas em um DW). Sem tais atualizações, seu conteúdo perde a "validade", não sendo mais tão confiável, tendo passado um tempo considerável no que diz respeito às abstrações em forma de regra que são extraídas dele. Outro aspecto interessante é o custo envolvido no projeto do *exploration warehouse*, que não pode ser justificado antes da obtenção de resultados; ao contrário de um projeto de DW, que poderá apresentar ferramentas analíticas que o utilizarão, um ambiente de exploração, além de temporário, conterá normalmente um conjunto de dados específicos demais para o uso generalizado de análise.

No tocante à alimentação do EW, subconjuntos de dados são movidos para ele a partir do *data warehouse* corporativo. A própria estrutura de tabelas do EW pode ser completamente modificada em relação ao DW, permitindo normalizações e agregações, transformando tuplas em registros, diminuindo a granularidade das informações, etc. Em resumo, o *exploration warehouse* é modelado e preenchido conforme for mais fácil e vantajoso para o objetivo da mineração de dados a ser aplicada.

3.2.6. O componente de armazenamento *near-line*

Segundo Inmon et al. (2001), neste tipo de armazenagem dados raramente utilizados são extraídos do *data warehouse* e gravados em dispositivos ópticos ou magnéticos. A idéia de armazenagem *near-line* contribui com a CIF nos seguintes aspectos:

1) diminuição do custo do DW em termos de hardware, visto que discos rígidos cujo valor é significativamente maior em relação a meios óticos e magnéticos irão destinar-se somente aos dados mais freqüentemente acessados;
2) ganho no desempenho de consultas ao reduzir a quantidade de dados total presente no DW; e
3) liberdade para inserção do mais baixo nível de granularidade de dados na medida em que o projetista tem espaço irrestrito para armazenagem de dados sem prejuízo de desempenho ou relevante custo de hardware.

3.2.7. A CIF e o Sistema de Suporte à Decisão (Decision Support System – DSS)

Uma fábrica de informações corporativas possui diversas vantagens para a análise e a extração de conhecimento. Mas a arquitetura completa de uma CIF (contendo todos os componentes descritos) não é necessária em todos os tipos de sistemas de informação organizacionais – na verdade a maioria das corporações não a possui. Também não é preciso criar seus componentes em paralelo, mas apenas à medida que forem sendo exigidos. É por isso que determinadas partes da CIF somente demonstram sua relevância e utilidade conforme o seu nível de maturidade.

Os componentes de uma CIF formam a base de todo o processamento dos Sistemas de Suporte à Decisão (INMON et al., 2001). O DSS é responsável por tornar os grandes volumes de dados incompreensíveis armazenados no DW em pequenas quantidades de informações de alta qualidade passíveis de entendimento pelos seres humanos (COLLARD et al., 2001).

A seguir, apresentam-se o conceito de *Data Mining* e o papel que as estruturas descritas até agora desempenham em suas muitas operações: de demanda de recursos computacionais; exigências de performance; e custos relacionados ao seu projeto.

3.3 *DATA MINING*

Mineração de dados é o processo de descoberta de conhecimento útil e relevante em grandes bases de dados. Informação útil, como sugerida por Singh et al. (2000), pode ser expressa pela relação entre proposições, entre outras coisas, que podem servir para predizer padrões e comportamentos futuros.

Para executar esse processo, são construídas aplicações que utilizam as mais diversas técnicas, buscando aumentar desempenho e confiabilidade, bem como permitir a otimização do processo de mineração de dados.

Assim, o *Data Mining* faz uso de técnicas para descobrir e apresentar conhecimento compreensível ao ser humano Frawley et al. (1992). Tem rapidamente envolvido áreas de pesquisa que fazem intersecção com outras disciplinas, incluindo estatística, banco de dados, reconhecimento de padrão (*PR – Pattern Recognition*), Inteligência Artificial, aprendizado de máquina, visualização e computação paralela de alta performance (FAYYAD, 1997).

Embora profissionais de várias áreas façam uso do termo "*Data Mining*", os primeiros a utilizá-lo foram os estatísticos (FAYYAD, 1997). Segundo Inmon et al. (2001), a mineração de dados não é uma área nova de estudos, pois já na década de 60 existiam pacotes de software para análises estatísticas de dados, cujas rotinas essenciais encontram-se no núcleo das principais ferramentas de DM atuais. Inicialmente, tais softwares foram utilizados no campo da Física e das Ciências Sociais, trabalhando-se com amostragens e rotinas analíticas. A facilidade de análise dos dados por cientistas leigos em processos estatísticos incentivou a demanda por ferramentas de mineração em outros setores.

Porém, de acordo com Piatetsky-Shapiro (2000), a primeira geração de sistemas para mineração de dados como a conhecemos hoje apareceu na década de 80, consistindo de ferramentas voltadas para pesquisas com foco em tarefas simples (construir classificadores, redes neurais, encontrar *clusters* em dados, visualização de dados, etc.). A segunda geração de sistemas de *data mining* surgiu em 1995 e preocupava-se com vários tipos de análises de dados, incluindo limpeza e integração. No final da década de 90, as necessidades dos usuários de negócio levaram ao aparecimento da terceira geração de aplicações e soluções baseadas em

data mining; dessa vez, orientadas a problemas específicos, impulsionando o uso de ferramentas de *front-end*.

3.3.1 Processo KDD

Segundo Fayyad (1997), a área de *data mining* focaliza-se apenas no escopo das técnicas e nos métodos de extração de conhecimento. Porém, o estágio em que a mineração de dados será aplicada requer que muitas outras atividades tenham sido desenvolvidas previamente. Na verdade, de acordo com Witten et al. (2000), a preparação dos dados para o uso de *data mining* consome a maior parte dos esforços investidos durante todo o processo. Cabena et al. (1998) chegam a estimar que a preparação dos dados engloba até 60% dos recursos destinados à aplicação de mineração de dados.

Então, para lidar com os vários aspectos e particularidades de cada ambiente de informação, preparando tal ambiente para a aplicação da mineração de dados, existe um processo maior e mais extenso do qual o *data mining* é apenas uma das atividades: o processo de Descoberta do Conhecimento (KDD – Knowlegde Discovery in Databases), definido por Fayyad (1996) como as atividades que abrangem desde a seleção dos dados até a análise dos resultados da mineração e a consolidação do conhecimento adquirido.

Assim, apesar de o termo *Data Mining* ser muitas vezes utilizado como sinônimo de KDD, Fayyad et al. (1996a) afirmam que DM é apenas um dos passos no processo de KDD. A mineração de dados encontra-se em um nível de abstração mais elevado, estando acima dos problemas específicos de cada organização, da maneira como as informações são administradas, da forma de armazenamento utilizada, das políticas de padronização e da entrada e saída de dados.

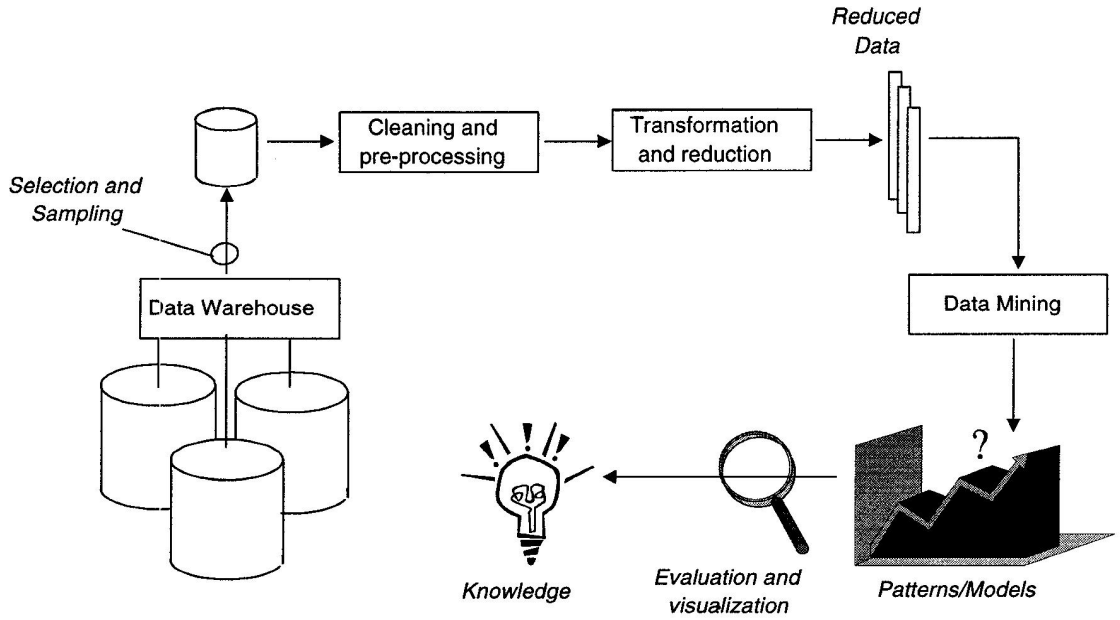

Figura 3.5 - Passos do processo KDD

FONTE: FAYYAD, 1997.

As tarefas de KDD citadas por Fayyad (1996) e apresentadas na Figura 3.5 são descritas a seguir.

- Entendimento do domínio de informação: focalizar o conhecimento que se deseja extrair através do processo.
- Pré-Processamento: escolher os atributos relevantes à análise, discretização e conversão de dados, ao tratamento de ruídos e valores ausentes, à transformação dos dados, etc..
- Restrição de dados: conforme o foco da análise a ser feita.
- Seleção da técnica de DM: escolher o método de mineração de acordo com o objetivo que a técnica possui (classificação, regressão, clusterização, etc.).
- Seleção do algoritmo mais adequado: baseando-se no problema, escolher o algoritmo ou processo computacional mais adequado para desempenhar a tarefa de DM objetivada.
- Aplicação da mineração: executar o algoritmo conforme a técnica e os métodos selecionados.
- Interpretação dos resultados: efetuar a análise heurística a partir dos resultados obtidos.

- Consolidação: validar o conhecimento encontrado através de indicadores ou da comparação com outros resultados atingidos através de outros métodos.

Embora os sistemas de banco de dados venham se desenvolvendo cada vez mais na direção da análise de informações, ainda restam muitos problemas quanto ao ambiente de informações, os quais estão mais relacionados à forma como os bancos de dados são utilizados pelas organizações do que propriamente com sua estrutura modelada. Sobre esses problemas, Matheus et al. (1993) levantam alguns muito freqüentes e comuns que desafiam o sucesso do processo KDD, e sugere soluções práticas para alguns deles:

- a dinâmica dos dados: as informações estão em constante mudança em um banco de dados, a validade de amostragens interfere na validade do conhecimento encontrado e por isso é essencial determinar corretamente os períodos em que a análise se aplica;
- ruído e incerteza: entradas de dados errados afetam a segurança do conhecimento encontrado e podem ser detectadas somente em grandes amostras, nas quais mais facilmente conseguem ser apontadas como *outliers*;
- dados incompletos: a ausência não somente de valores em certos campos dos registros mas também de campos de dados necessários para a análise (falha no projeto de banco de dados) impede avaliações e explorações realmente abrangentes;
- redundância de informação: dados transformados, agregados ou com dois tipos de unidade diferentes (por exemplo, bruto e em porcentagem) causam dependências herdadas, prejudicando as análises (falsa correlação natural, função de regressão induzida, etc.);
- dados esparsos: os eventos de interesse da exploração podem representar uma quantidade insignificante de registros na base, resultando em amostras inválidas para um robusto processo de reconhecimento de padrões;
- volume de dados: o enorme volume de informações obriga que a análise seja feita em amostras, randomicamente selecionadas ou restritas a subclasses de registros possivelmente mais relevantes à exploração; e

– sumarização dos dados: focalizar o domínio de informação a ser pesquisado é importante para aumentar a qualidade da base de exploração, permitindo assim maior refinamento e limpeza dos dados.

Alguns dos pontos anteriormente mencionados poderão ser resolvidos pelas atividades de KDD precedentes à aplicação de mineração. Mas há certos problemas levantados que somente podem ser identificados e tratados tendo-se em mente qual o propósito da exploração, por exemplo, os dados esparsos e a sumarização.

Os diferentes ambientes de informação e o variado conhecimento existente neles indiretamente geram distinção entre as técnicas de *data mining* selecionadas pelo analista. Embora não obriguem a aplicação de uma metodologia específica, problemas que demandam as mais comuns tarefas executadas por aplicações de DM – busca por reconhecimento de padrões, resultando em atividades de clusterização, classificação, regras de associação, análise de regressão, etc. (COLLARD et al., 2001) – têm sido largamente explorados em recentes trabalhos científicos. Os resultados apontam que algumas abordagens de mineração são mais efetivas do que outras no tocante a determinadas demandas de conhecimento.

3.3.2 Processo indutivo ou intuitivo, dedutivo e analítico

É possível definir a etapa em que se encontra a atividade de extração de conhecimento de acordo com a forma de visualizar o conhecimento através do processo de obtenção do mesmo. Durante a exploração, parte-se de uma intuição para criar uma hipótese, ou seja, o analista é induzido pelo senso comum – no conhecimento prévio de um ambiente –, produzindo uma assertiva. A dedução sobre a validade da hipótese é obtida através da mineração de dados. O uso de metodologia formal para comprovar e interpretar a hipótese torna o processo analítico.

3.3.3 O processo de exploração

A exploração é a atividade do processo de *Data Mining* em que as hipóteses são geradas. Ao contrário do que o termo parece definir, a exploração e a mineração são dois

diferentes conceitos, os quais são claramente definidos por Inmon et al. (2001), seu criador, da seguinte forma:

> A exploração admite que existem dados que podem estar escondendo alguns padrões de comportamento interessantes e úteis. A partir desses padrões, as hipóteses são traçadas. A mineração de dados assume somente que há uma hipótese; seu propósito é testar a validade e a força da hipótese. A exploração precisa que a mineração de dados teste a hipótese que foi descoberta. A mineração de dados necessita que a exploração identifique a hipótese a ser testada. (INMON et al., 2001, p. 3).

Embora os termos estejam intimamente relacionados, ainda segundo Inmon et al. (2001), ambos os processos existem em complemento mútuo para formar uma entidade holística. Durante a exploração, o "explorador" representa o pensador da corporação, observando aspectos ainda não percebidos; porém, tanto o explorador (analista técnico) como o empresário (analista de negócios) precisam trabalhar em conjunto para que o processo de exploração seja ao mesmo tempo eficiente sem deixar de ser relevante para os negócios.

Na busca de simples padrões, relações entre os dados, ocorrências incomuns, etc., as técnicas estatísticas são ferramentas significativas para o explorador. São elas que determinam a correta estratificação de amostras, detectam *outliers* através de análises como a do desvio-padrão, encontram dependências pela correlação, efetuam discretizações, executam a avaliação das medidas básicas da população de dados (mediana, distribuição pela Normal, resíduos e R^2 em análise de regressão, etc.), entre outras importantes tarefas.

Quanto a essas técnicas, é necessário lembrar que a Mineração de Dados de maneira alguma substitui os já conhecidos métodos estatísticos ou de aprendizado de máquina, mas apenas estende essas áreas para aplicações em grandes bases de dados (COLLARD et al., 2001).

A exploração, de acordo com Inmon et al. (2001), é um processo heurístico ou uma análise repetitiva. O autor estabelece tal conceituação ressaltando uma fundamental diferença em relação aos demais tipos de processamento analítico, explicando que na análise heurística cada passo é planejado e executado conforme os resultados obtidos no passo anterior, ou seja, na análise heurística as idéias iniciais vão sendo refinadas com o andamento do próprio processo analítico, ao contrário dos outros tipos de análise, que são projetadas e especificadas mesmo antes que qualquer trabalho seja desenvolvido.

Essa definição explica como basicamente desenvolvem-se as atividades de exploração de dados: uma idéia sobre um conhecimento potencial e ainda oculto é refinada a partir de um processo iterativo, cujos resultados são avaliados a cada passo para redirecionar e até mesmo definir as próximas atividades de análise.

Mas, para que o explorador construa uma base de dados de qualidade voltada ao propósito de sua exploração e modelada dentro de uma estrutura de BD que dê suporte à carga de processamento que demandam suas consultas, o analista deve considerar alguns aspectos importantes já citados e detalhados a seguir. É importante que essa base fique disponível pelo tempo de que necessitarem suas tarefas de busca, entre outras atividades do processo.

3.3.4 O processo de amostragem

A qualidade da base de dados começa com a seleção de dados relevantes para a análise. Uma análise de padrões que utilize o nível de confiança mais comumente aplicado implicaria em se trabalhar com uma variedade de registros em um intervalo aproximado entre 5% e 95% da base de dados. No entanto, para certas aplicações de manipulação analítica de dados, o processamento computacional de tamanha base de informações (considerando grandes BDs organizacionais e o extremo de 95%) se torna inviável se considerarmos os custos envolvidos em hardware e software.

A resposta para esse problema é conhecida: trata-se da seleção de amostras randômicas ou direcionadas às classes de informação de interesse do analista. Muitos estudos na área de Estatística têm sido dedicados ao processo de coleta de amostras. Os métodos desenvolvidos para lidar com essa tarefa, envolvida em quase todos os tipos de análise estatística, são conhecidos como Técnicas de Amostragem.

Durante o processo de amostragem o analista avalia a quantidade satisfatória de registros disponíveis à exploração desejada, os conjuntos de informação em que focalizará a geração de hipóteses, a granularidade e a sumarização realmente necessárias para se pôr em prática a idéia inicial já em mente.

3.3.5 Detecção de outliers

A detecção de *outliers* é realizada por várias técnicas de DM, como a análise de regressão (através da medida de dispersão dos dados), as redes neurais (por análise de *clusters*) e até mesmo pelo número de desvios padrão apresentado no conjunto de dados.

Como bem observado por Williams et al. (2002), existem muitas aplicações em que a descoberta de *outliers* torna-se mais interessante do que a análise dos indivíduos dentro do padrão encontrado, como, por exemplo, as aplicações de mineração para detecção de fraudes ou para análise de limites superiores e inferiores de uma população de dados – os maiores consumidores, os menores valores para os índices de queda de tensão elétrica, etc.

A Estatística identifica três categorias de *outliers*:

1) *cluster*: são dados separados do conjunto principal e reunidos em um *cluster* com pequena variância interna;
2) radial: ocorrem nas bordas-limites do grupo principal de dados; e
3) disperso: pontos randomicamente espalhados ao longo da maior concentração de dados.

A proporção de *outliers* em uma população ou amostra é chamada de “grau de contaminação” do conjunto de dados (WILLIAMS et al., 2002). Há diferentes avaliações quanto a esse grau, sendo que a Estatística pode considerar níveis de 40% ou mais, enquanto a Mineração de Dados só se importa com valores bem menores, geralmente inferiores a 4%. Isso se explica pelo fato de que não é desejável, na busca de informação rara, que mais de 40% do banco de dados seja constituído de exceções. Neste trabalho, será utilizado o conceito de DM para avaliar *outliers*.

3.3.6 Armazenagem dos dados de exploração

De acordo com Inmon et al. (2001), o lugar ideal para executar a exploração é o *exploration warehouse*, no qual o analista poderá construir sua própria estrutura relacional (entre outras), carregar somente os dados de seu interesse a partir do *data warehouse*

corporativo e dispor de um ambiente exclusivo para as atividades pesadas de análise. Porém, tanto o *data warehouse* como a armazenagem *near-line* também são fontes apropriadas em termos de qualidade de dados para as operações de *Data Mining*.

3.3.7 Validade temporal dos dados

Embora o reconhecimento de padrões inerentemente exija dados históricos, as fronteiras para o intervalo de tempo coletado baseiam-se muito na intuição e na experiência do analista. Paralelamente ao julgamento efetuado pelo explorador, as orientações do analista de negócio são importantes na medida em que manterão o técnico direcionado à obtenção do conhecimento que é esperado no final do processo (por exemplo, determinar a sazonalidade nas vendas de um produto requer que a mesma época do ano seja comparada ao longo de alguns anos; uma amostra englobando apenas diferentes épocas não teria utilidade para extrair essa hipótese).

3.3.8 Reutilização das amostras

Diretamente relacionada à política de armazenagem e temporalidade dos dados está a reutilização das amostras para novas análises. Ao fazer uso de sistemas analíticos, Matheus et al. (1993) afirma que a armazenagem e a reutilização das descobertas até então feitas sobre os dados são importantes para que tais sistemas aprendam com as experiências realizadas.

No tocante ao uso do *exploration warehouse* para armazenagem, quatro tipos diferentes dele exemplificam as possíveis abordagens aplicadas à reutilização da amostra de exploração:

- estáticos e temporários;
- estáticos e permanentes;
- dinâmicos e temporários;
- dinâmicos e permanentes.

Estático ou dinâmico refere-se à freqüência de atualização do *exploration warehouse*. Temporário ou permanente diz respeito ao período de tempo em que a estrutura será utilizada.

EWs estáticos e temporários bem como EWs dinâmicos e permanentes são os mais encontrados na prática.

Observa-se ainda que a correta e clara documentação sobre o *exploration warehouse* e seu conteúdo são imprescindíveis para que se possa reutilizá-lo seguramente em novos processos de análise.

3.3.9 *Data Mining* e o reconhecimento de padrões

O próprio conceito de Mineração de Dados regularmente se confunde e se mistura com a atividade de descoberta de padrões (GORODETSKY, 2003). Porém, a relação entre as áreas é a de uma intersecção em que nenhum dos dois domínios de aplicação, métodos e características abrange o outro. De acordo com Duda (1973):

> O reconhecimento de padrões é um campo que se preocupa com o reconhecimento por máquina de regularidades significativas em ambientes com ruído ou complexos.

O autor ainda afirma que não há uma teoria simples de reconhecimento de padrão que consiga abranger todos os tópicos importantes devido à singularidade de cada domínio de aplicação.

Conforme Schalkoff (1992), o PR caracteriza-se como um processo de redução, mapeamento ou rotulação da informação. Este autor destaca a diferença entre o conceito de característica (*feature*) e padrão (*pattern*): padrão pode ser simplesmente um conjunto de medidas ou observações representadas em vetores ou matrizes; já característica é qualquer medida de extração utilizada.

Existem três abordagens principais para o reconhecimento de padrões:

1) estatística (ou teórica de decisão);
2) sintática (ou estrutural); e
3) neural.

Do mesmo modo como o DM é uma área multidisciplinar, as técnicas de PR relacionam-se com outras áreas de conhecimento (DUDA, 1973), entre as quais estão os sistemas de processamento de sinais (adaptativos), a inteligência artificial, a modelagem neural, a teoria da comunicação, os conjuntos difusos, a psicologia, a teoria de autômatos, a teoria de controle e as linguagens formais (lingüística).

Mas alguns pontos essenciais distinguem essas duas áreas de conhecimento, e a principal diferença está nos conceitos de descobrir e reconhecer. Basicamente o PR não descobre padrões, apenas os reconhece, isto é, identifica padrões já conhecidos, sendo uma de suas maiores aplicações a Classificação (DUDA, 1973). Enquanto isso, a mineração de dados, nesse contexto, interessa-se somente pela descoberta de novos padrões e por sua validação. Em PR, os padrões já estão validados.

Indo mais além, verifica-se que a área de reconhecimento de padrões possui a capacidade de extrair características de um objeto, transformá-las em dados e classificar o objeto segundo padrões já conhecidos (por exemplo, identificação de impressões digitais e análise de texturas). A atividade de mineração obrigatoriamente parte do princípio da existência de dados, deixando para as tarefas de KDD (anteriores a ela) toda a extração e preparação desses dados. Além disso, o DM executa a busca orientando-a a um foco de forma a considerar o interesse da análise, ou seja, nem todo padrão encontrado constitui-se em conhecimento não óbvio ou útil.

Como afirmado por Matheus et al. (1993), a combinação de novos domínios de conhecimento e técnicas empíricas deverá se tornar cada vez mais importante para o processo de reconhecimento de padrões em DM, visto que as pessoas estarão buscando descobrir não somente qual o padrão mas também o porquê de sua ocorrência entre os dados.

Para que o processo seja efetivo, algumas condições básicas devem ser seguidas conforme Inmon et al. (2001), tais como o nível de detalhe adequado e as diversas ocorrências das variáveis múltiplas e com dados que possuam certa homogeneidade.

3.3.9.1 Relação entre as variáveis e a análise de correlação

Determinar a relação existente entre as variáveis (campos valorados) de um conjunto de informações é importante para que se possa definir a causalidade dos padrões encontrados, além de definir a força com que essas variáveis agem sobre o comportamento da outra.

Inmon et al. (2001) ressaltam a necessidade de se observarem a força da relação encontrada, sua natureza e a inter-relação entre os fatores causais, identificando três tipos de relação possível entre as variáveis:

1) relação causal direta: é a mais forte, mais simples e mais rara de ser encontrada;
2) relação indireta: também chamada correlativa, é a mais comum, porém pode ser complexa;
3) relação randômica: relação em que não há um padrão de comportamento identificável entre as variáveis.

A medida de correlação estatística irá ajudar a definir a força das relações existentes no conjunto de informações, permitindo descartar variáveis que não estão envolvidas com o ponto de interesse focalizado pelo analista ou fazendo-o perceber pontos anteriormente tidos como irrelevantes do ambiente de dados.

3.3.9.2 Análise de tendência

Considerando que os dois principais objetivos de DM são a descrição e a predição (COLLARD et al., 2001), a análise de tendência torna-se uma das tarefas mais comumente encontradas no processo de mineração.

A análise de tendência não é necessariamente feita sobre um eixo temporal. Qualquer intervalo de valores de uma variável, devidamente valorada para todos os dados da análise, permite gerar uma função matemática que demonstra o comportamento aproximado da informação ao longo dessa variável (por exemplo, análise do índice de carregamento de potência pela quantidade de consumidores ligados àquele circuito elétrico em que o número de consumidores está em uma escala que varia de 10 até 100).

3.3.10 Técnicas de *Data Mining* utilizando Inteligência Artificial

Quanto às técnicas de *Data Mining*, revisa-se neste trabalho apenas aquelas que são passíveis de serem aplicadas no contexto do problema estudado, com especial ênfase nas técnicas de computação evolucionária. De acordo com Berry (1997) as técnicas mais utilizadas são:

1) análise de seleção estatística: corresponde a agrupamentos para descrição de conjuntos que tendem a ocorrer através de probabilidade. Esses agrupamentos podem ser expressos como regras;
2) MBR (Memory-Based Reasoning): utiliza-se de exemplos anteriores conhecidos para fazer previsões de exemplos desconhecidos;
3) agrupamentos: objetiva encontrar padrões de dados semelhantes para que estes sejam agrupados em classes;
4) árvores de decisão: divide registros do conjunto de dados em subconjuntos, produzindo regras associadas a um ou mais campos da base;
5) redes neurais: constituem modelos que imitam as interconexões no cérebro. Aprendem com um conjunto de dados de treinamento e são geralmente utilizadas para classificação ou previsão;
6) algoritmos genéticos: baseiam-se nos conceitos da genética e seleção natural. Utilizam operadores tais como seleção, *crossover* e mutação para realizar buscas em uma base de dados.

Algoritmos genéticos também podem ser aplicados em conjunto com redes neurais artificiais na determinação dos pesos e da sua arquitetura. Além disso, podem ser utilizados para a geração de regras de produção ou listas de decisão. Quando uma técnica é aplicada em conjunto com outra, a método resultante é chamado de “híbrido”.

Na busca de improver a capacidade de compreensão do usuário ao conhecimento resultante das técnicas de *data mining*, a abordagem difusa pode ser aplicada à mineração devido a sua flexibilidade e seu poder para tratar incertezas além ser um meio natural para representar regras com atributos contínuos (FERTIG et al., 1999). Conjuntos Difusos e Lógica Difusa são uma alternativa para a lógica booleana (Zadeh, 1965) que determina um domínio dentro de um intervalo com somente dois pontos geralmente extremos (verdadeiro e falso,

certo e errado, quente ou frio, etc.). A abordagem difusa, ao contrário, permite a codificação direta de conhecimento através da formação de uma descrição linguística difusa (TSOUKALAS & UHRIG, 1997).

A Teoria dos Conjuntos Difusos forma a base para a Lógica Difusa, permitindo a construção de expressões lógicas e raciocínios aproximados (ROMÃO et al., 2002). Ao utilizar palavras de linguagem natural, a variável mapeada dentro de um conjunto difuso é chamada de "variável lingüística" e os conjuntos difusos determinados para essa variável são denominados "termos linguísticos". Desse modo, o valor de uma variável indicando temperatura, por exemplo, pode ser caracterizado quanto a sua pertinência a um extremo do intervalo de seu domínio, isto é, em vez de caracterizá-lo somente entre "quente" ou "frio", pode-se classificá-lo dentro dos termos linguísticos "pouco quente", "muito frio", etc.

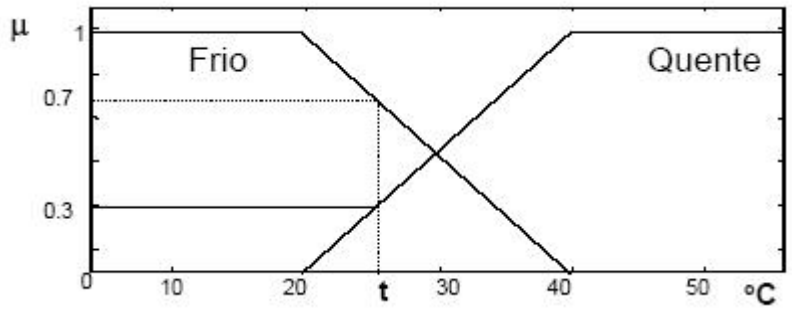

Figura 3.6 - Conjuntos difusos de temperatura

FONTE: ROMÃO, 2002.

Para fazer tal classificação são usadas funções de pertinência (μ) - FPs. As formas mais empregadas de função de pertinência segundo Romão (2002) são a trapezoidal (apresentada na Figura 3.6 com o exemplo de uma variável de temperatura) a triangular e a gaussiana.

Em geral o valor máximo de uma FP é 1, pois é uma medida relativa da aproximação de um valor ao seu máximo ou mínimo possível em relação ao seu escopo inteiro. Por exemplo, a luminosidade de um ambiente está em 0,4 escuro e 0,6 claro ou em 40% escuro e 60% claro; onde 0 (conjunto difuso escuro) é a luminosidade mínima existente e 1 (conjunto difuso claro) é a luminosidade máxima possível.

Fatos difusos são representados por regras difusas. Nesse caso, a determinação do grau de pertinência do antecedente de uma regra é chamada de “fuzzificação”. O processo de se extrair o valor mais característico (típico) de um conjunto difuso se chama “defuzzificação”.

3.3.11 Regras como representação dos resultados

Neste trabalho a extração de regras se destaca como um dos objetivos principais do estudo. Por isso, é importante que a tarefa não seja confundida com a forma de representação, pois para executar a “tarefa” de extração de regras existem diferentes “métodos”, os quais, por sua vez, possuem muitas maneiras de representar seus resultados (conjuntos, matrizes, árvores, gráficos, etc.).

Assim, é necessário ressaltar-se também que a regra em si é apenas uma forma de exibir os resultados obtidos pela técnica. Diferentes tarefas utilizam o formato de regras – por exemplo as tarefas de Associação, Classificação e Regressão –, pois se trata de um formato bastante compreensível ao ser humano. Regras comumente são do tipo “SE... ENTÃO” (conhecidas como regras de produção. Em geral regras contêm um antecedente (ou premissa) e um conseqüente. Segundo Romão (2002):

- o *antecedente* é formado por expressões de condição contendo atributos do banco de dados; e
- o *conseqüente* é formado por uma expressão indicando a previsão de um atributo meta como resultado do conjunto de atributos da premissa (antecedente).

Considerando U como sendo todo o conjunto de atributos na base de dados, a representação de uma regra de associação, como descrita em Richards et al. (2001), pode ser feita da seguinte forma:

$$\textbf{\textit{antecedente}} \Rightarrow \textbf{\textit{conseqüente}}$$

onde:

- antecedente $\subset U$
- conseqüente $\subset U$
- antecedente $\cap$ conseqüente = $\varnothing$

– $\Rightarrow$ pode significar construção, conjunção, disjunção, igualdade, diferença, entre outros operadores.

Isto é, tanto o *antecedente* como o *conseqüente* são subconjuntos de *U*, contendo atributos do banco de dados, mas não há interceção de atributos entre ambos.

Quanto ao significado das regras, segundo Liu & Hsu (1996) as regras podem diferir sendo quanto ao antecedente inesperado, antecedente contraditório e conseqüente contraditório. Ou seja, uma regra pode diferir de uma hipótese (ou regra previamente encontrada) em dois sentidos diferentes: conseqüentes e/ou condições. As distinções se mostram através dos atributos que compõem as partes e também nos valores que eles possuem. Considerando o seguinte exemplo de hipótese:

Hipótese EX: (Período = "manhã"), (Causa = "sobrecarga") => DEC = "alto"

a) Conseqüente inesperado: a parte condicional da regra é igual à da hipótese e o conseqüente de ambos possui o mesmo atributo, mas os valores do conseqüente são diferentes. Exemplo:

Regra CqI: (Período = "manhã"), (Causa = "sobrecarga") => DEC = "médio"

b) Conseqüente contraditório: o antecedente da regra é igual ao da hipótese, mas o atributo é distinto nos conseqüentes. Exemplo:

Regra CqC: (Período = "manhã"), (Causa = "sobrecarga") => Mês = "baixo"

c) Condição inesperada: o atributo do conseqüente na hipótese e na regra possui o mesmo valor, porém os valores condicionais no antecedente da regra são distintos. Exemplo:

Regra CdI: (Período = "tarde"), (Causa = "programada") => DEC = "alto"

d) Condição contraditória: o antecedente da regra possui atributos diferentes do antecedente da hipótese, embora atributos e valores no conseqüente de ambos sejam os mesmos. Exemplo:

Regra CdC: (Mês = "médio"), (Potência Nominal = "baixo") => DEC = "alto"

3.3.12 Tarefas comuns realizadas por *Data Mining*

O conjunto de técnicas e ferramentas de DM é selecionado conforme o tipo de tarefa a ser realizada. Utilizar uma técnica de mineração de dados às cegas, sem analisar a sua adequação para uma data tarefa, é um dos maiores enganos cometidos pelos analistas (WITTEN et al., 2000). Os objetivos mais comuns para se utilizar mineração de dados são descritos brevemente a seguir.

3.3.12.1 Clusterização

A clusterização ou segmentação é uma técnica que agrupa um conjunto de dados, maximizando as similaridades entre os dados dentro do mesmo *cluster* e minimizando as similaridades entre *clusters* diferentes (GARAI et al., 2003). O objetivo principal é definir quais e quantos conjuntos agrupados por características semelhantes (padrões) existem na base de dados, automaticamente gerando a descrição das classes encontradas. Os métodos de clusterização podem ser classificados como hierárquicos e não hierárquicos. Entre os métodos não hierárquicos, o algoritmo mais conhecido para clusterização é o *k-means* (GARAI et al., 2003).

Algoritmos de clusterização tipicamente dividem-se em dois estágios: um laço externo para trabalhar o número de possíveis *clusters* e um laço interno para adequar a melhor clusterização a um determinado número de *clusters*. Quando um número qualquer de *clusters* é dado, os métodos dividem-se em três tipos: (1) baseados na métrica da distância (*metric-distance based*); (2) baseados no modelo (*model-based*); e (3) baseados em partições (*partition-based*) (FAYYAD, 1997).

Além de encontrar *clusters* e associar os dados a eles, o desafio dessa tarefa de mineração é conseguir associar novas instâncias de dados aos *clusters* já existentes (WITTEN et al., 2000).

3.3.12.2 Modelo de previsão

Segundo Fayyad (1997), se o campo a ser predito possui valor contínuo, então se trata de uma análise de regressão; porém, se for uma variável categórica, o problema é de classificação. Quanto à regressão, linear ou não-linear, a transformação dos dados de entrada é uma dificuldade que requer conhecimento das regras de negócio; em relação à classificação, a transformação também é às vezes chamada de "extração de característica" (*feature extraction*).

Com relação a problemas de predição numérica, em que a saída a ser predita não é uma classe discreta, mas uma quantidade numérica, saber quais são os atributos importantes e sua relação com o valor de saída é mais importante que a tarefa de predizer valores para novas instâncias (WITTEN et al., 2000).

3.3.12.3 Associação

A busca por regras de associação, também chamada ARM (*Association Rule Mining*), é uma das tarefas mais populares, sendo resumidamente definida como a descoberta de pares de conjuntos de elementos que tendem a aparecer juntos em determinados contextos (SCHUSTER, 2003). O algoritmo mais conhecido para executar ARM é o *Apriori* (AGRAWAL, 1994), embora vários algoritmos tenham sido desenvolvidos para extrair regras de associação (para revisão dos algoritmos, ver Schuster (2003)).

Em sua forma básica, Agrawal et al. (1993) afirmam que a descoberta de regras de associação é uma tarefa determinística, em que não ocorre previsão. Os autores propuseram ainda um modelo matemático para avaliar a relevância da regra encontrada, considerando o suporte e a confiança da regra. "Suporte" é a freqüência com que o antecedente da regra ocorre na base de dados; já "confiança" é o número de vezes em que o conseqüente da regra se apresenta juntamente com o antecedente.

3.3.12.4 Classificação

A classificação é a tarefa de associar registros a classes predefinidas, descobrindo relacionamentos entre os registros através de seus atributos (HAND, 1997). Para predizer o estado de semelhança de uma variável categórica é comumente usada a medida de estimação de densidade, que inclui as técnicas de estimação de densidade, a métrica espacial e a projeção em regiões de decisão (FAYYAD, 1997). O aprendizado por meio de classificação pode ser chamado de supervisionado visto que a saída esperada (a classe) é informada pelo usuário.

A definição estatística de “classificação” é bem similar, sendo descrita como a atividade de associar os dados de entrada a uma ou mais classes pré-especificadas de acordo com a extração de características ou atributos significantes e com o processamento ou a análise desses atributos (SCHALKOFF, 1992). Na literatura estatística, esse tipo de aprendizado comumente é referenciado como “discriminação” (ROMÃO, 2002).

Uma das saídas mais comuns produzidas pela tarefa de classificar é no formato de regras. Existem importantes diferenças entre as Regras de Associação e as Regras de Classificação (ou de Previsão). A atividade de associação busca representar padrões e regularidades, caracterizando os dados, enquanto a classificação distingue os dados de acordo com os seus aspectos e os associa a uma classe. Conforme Witten et al. (2000), a associação difere em dois principais pontos: (1) pode predizer qualquer atributo, não apenas a classe; e (2) pode predizer o valor de mais de um atributo ao mesmo tempo.

Há muitas técnicas de *Data Mining* que podem ser utilizadas para efetuar a tarefa de classificação, dependendo principalmente do tipo de conjunto de dados (KING, 1995) disponível. A Figura 3.7 apresenta as mais conhecidas.

Figura 3.7 - Técnicas de *Data Mining* utilizadas para a tarefa de classificação

FONTE: adaptado de: ROMÃO, 2002.

O sucesso da atividade de classificação pode ser medido submetendo-se a descrição conceitual aprendida a conjuntos de dados para teste. Quanto maior for a taxa de acerto da descrição gerada aplicada aos dados de teste ou quanto mais aceitável for a descrição para o usuário, mais bem aprendido foi o conceito (WITTEN et al., 2000).

Segundo Fayyad et al. (1996a), o primeiro passo deve ser decidir a tarefa de mineração de dados necessária à aplicação. A escolha da técnica está essencialmente ligada ao negócio, à aplicação e à quantidade de dados que estão à disposição do analista (GONÇALVES, 2000). A Tabela 3.2 relaciona as principais técnicas selecionadas conforme a tarefa de DM escolhida e de acordo com as mais comuns aplicações de mineração.

Funções	Algoritmos	Aplicações
Associação	Estatística Teoria dos conjuntos	Análise de mercado
Classificação	Árvores de decisão Redes neurais Algoritmos genéticos	Controle de qualidade Avaliação de riscos
Agrupamentos	Redes neurais Estatística	Segmentação de mercado
Previsão de séries temporais	Estatística Redes neurais	Previsão de vendas Controle de estoque
Padrões seqüenciais	Estatística Teoria de conjuntos	Análise de mercado ao longo do tempo

Tabela 3.2 - Relação das tarefas, técnicas e aplicações de Mineração de Dados

FONTE: BIGUS, 1996.

3.4 CONSIDERAÇÕES FINAIS

Este capítulo revisou definições de *Data Warehouse* e *Data Mining*, dando ênfase ao uso de técnicas de Mineração de Dados aplicadas em conjunto com Inteligência Artificial. Grande parte dos conceitos citados é utilizada posteriormente neste trabalho no processo de montagem do ambiente para exploração de informação, bem como dá suporte e guia a definição e a seqüência dos procedimentos para extração de conhecimento. O próximo capítulo introduz Algoritmos Genéticos, suas principais características, seus conceitos, seu funcionamento e sua aplicação em geral.

4 ALGORITMOS GENÉTICOS

4.1 HISTÓRICO

Os estudos sobre Computação Evolucionária começaram nas décadas de 50 e 60. Inicialmente, a idéia era aplicar a evolução para otimizar problemas de engenharia, utilizando operadores inspirados na seleção natural e na variação genética para evoluir uma população de possíveis soluções a um dado problema (MITCHELL, 1996). A programação evolucionária surgiu em 1966, com Fogel et al. (1996).

Os Algoritmos Genéticos foram inventados por John Holland na década de 60. Seu trabalho foi desenvolvido posteriormente em conjunto com colegas e alunos na Universidade de Michigan. Ao contrário dos estudos sobre estratégias evolucionárias e programação evolucionária, a idéia inicial de Holland era formalmente estudar o fenômeno de adaptação natural e desenvolver meios para que esses mecanismos pudessem ser importados para sistemas computacionais (MITCHELL, 1996).

Muitos trabalhos de Holland realizados na década de 60 demonstram o seu interesse em sistemas adaptativos (HOLLAND, 1962, 1965, 1966), reconhecimento de padrões (HOLLAND, 1969) e adaptação paralela (HOLLAND, 1973). No entanto, foi a publicação de seu livro, em 1975, que apresentou os AGs como a abstração da evolução biológica, introduzindo o uso de um algoritmo capaz de aplicar a simulação dos operadores naturais de *crossover*, inversão e mutação sobre uma estrutura em forma de população (HOLLAND, 1975).

Entre os trabalhos do grupo de estudo de Holland (1986), encontra-se a definição de uma abordagem básica para a utilização de AGs na tarefa de classificação. A abordagem Michigan recebeu o nome da universidade onde foi desenvolvida. Também conhecida como *Classifier Systems*, essa abordagem define que cada regra de classificação é representada por um indivíduo. Assim, a solução do problema é representada por um conjunto de regras ou de indivíduos (ROMÃO, 2002).

Mitchell (1996) aponta três razões fundamentais para o uso de mecanismos evolucionários na solução de problemas computacionais, as quais são apresentadas a seguir.

1) *Paralelismo*. Muitos problemas exigem a busca da solução através de um grande número de possibilidades. Por meio do paralelismo, várias soluções são exploradas simultaneamente de maneira eficiente, isto é, existe uma estratégia inteligente para selecionar o próximo conjunto de indivíduos a ser avaliado.
2) *Adaptação*. É muito comum encontrar problemas que demandem do software estabilidade em desempenhar seus objetivos, independente das mudanças em seu ambiente – robótica e outros sistemas com inteligência artificial possuem essa característica intrínseca. Adaptar-se ao meio, competindo ou cooperando, é justamente o aspecto mais interessante das técnicas evolucionárias, pois representa a capacidade da espécie de aprender ao longo do tempo de acordo com o ambiente.
3) *Inovação*. É a habilidade de gerar soluções completamente novas, não presentes no espaço conhecido de busca. Também quanto a esse ponto, a evolução natural demonstra enorme poder para gerar conteúdo genético novo, ou seja, produzir indivíduos que não estejam presentes na atual população.

4.2 TERMINOLOGIA

A Computação Evolucionária, ao implementar processos computacionais que imitam os processos de evolução, acaba por utilizar-se de vários termos da Biologia, mais especificamente da Genética. Seguem alguns desses conceitos empregados por Abercrombie et al. (1970) e Aurélio (1999), acompanhados dos respectivos termos na Computação Evolucionária (SCHNEIDER, 1998; MITCHELL, 1996), para melhorar a compreensão do paralelo natural feito pela disciplina.

1) Cromossomos: são cadeias de DNA (ADN – Ácido Desoxirribonucléico) constituídas por genes. AG: *strings*.
2) Genes: correspondem às características possíveis de aparecerem em um indivíduo, podendo estar ativos ou inativos. AG: característica, aspecto, locus na *string*.

3) Alelos: são os valores contidos em cada gene. AG: valor da característica.
4) Genoma: representa todo o material genético de um indivíduo. AG: solução completa.
5) Genótipo: diz respeito ao conjunto de genes contidos no genoma. AG: estrutura, cromossomos codificados.
6) Fenótipo: são as características observáveis, visíveis, de um indivíduo. AG: conjunto de parâmetros, solução alternativa, estrutura decodificada.
7) Indivíduo: é um exemplar de uma espécie que interage com o meio ambiente. AG: o mesmo que o cromossomo.
8) Haplóide: são seres cujo cromossomo não possui respectivo par.
9) Diplóide: correspondem às espécies que possuem um par de cada cromossomo em células somáticas (não sexuais).
10) *Fitness*: definida como a probabilidade de o organismo viver para se reproduzir, representa a sua adequação ao ambiente, de adaptação segundo um critério. AG: também chamada de função de *payoff* ou função objetivo.
11) Espécie: caracterizam-se por grupos de indivíduos capazes de se cruzar que são isolados reprodutivamente de outros grupos semelhantes, contendo fenótipos semelhantes. AG: indivíduos componentes de uma mesma população.
12) Seleção natural: trata-se de um processo que garante aos indivíduos mais aptos chances maiores de reprodução. AG: determinada pela aptidão do indivíduo, representa as chances de ele gerar descendência.
13) Adaptabilidade: refere-se a qualquer característica de um organismo vivo que aumenta as possibilidades de sobrevivência e de deixar descendência no seu ambiente. AG: qualquer alteração na estrutura de um cromossomo que melhore a sua capacidade de resolver determinado problema, permitindo-lhe sobreviver e se reproduzir mais.

4.3 SCHEMA E HIPERPLANO

A noção de *schema* (no plural: *schemata*) foi também introduzida por Holland (1968, 1975). Para Goldberg (1989), a chave para a abordagem com algoritmos genéticos é a construção de blocos de hipóteses – *building blocks* –, combinações de valores que conferem

alto *fitness* às *strings* (séries) nas quais estão presentes (MITCHELL, 1996), que parte do conceito de modelo de similaridade (*similarity template*) ou *schema*. A idéia principal é que a população de *strings* pode prover informações para direcionar a busca, melhorando o seu desempenho.

Um *schema* é um modelo de similaridade constituído de um conjunto de *strings* com similaridades em certas posições do *schema*. Um modelo é formado pelo alfabeto {0, 1, *}, ou seja, por números um (1), zeros (0) e asteriscos (*), sendo o asterisco um bit ainda desconhecido que pode significar tanto o número um (1) quanto o número zero (0). Quanto mais bits conhecidos houver no *schema*, menor será o espaço de busca (que se resumirá às possíveis combinações de valores para as posições contendo asteriscos naquele *schema*) e, conseqüentemente, maior será o desempenho. E, por se tratar de uma característica própria de AG e por possuir tanta influência no processamento, Goldberg (1989) chamou esse aspecto de paralelismo implícito (*implicit parallelism*).

Pode-se visualizar o processamento de um *schema* de três formas diferentes: (1) usando-se a própria visualização do *schema;* (2) através do problema do menor erro (*deceptive*); e (3) por meio de uma representação geométrica. O conceito de “espaço de busca” é mais facilmente compreendido pelo ser humano quando se utiliza a representação geométrica. Basta imaginar um espaço n-dimensional (onde *n* é o tamanho do cromossomo); cada um dos planos é o escopo possível de cada gene; e o alelo de um gene é um ponto nesse plano. O conjunto de pontos forma um *schema* preenchido (ou definido).

O espaço formado pelos possíveis valores para os genes que compõem o cromossomo é chamado de hiperplano. À medida que o tamanho do cromossomo cresce apenas a representação gráfica fica mais difícil de ser feita. Mesmo assim, a figura dimensional nos ajuda a compreender melhor o seguinte: quando o AG é guiado, torna-se desnecessário percorrer todos os planos do hiperplano, mas tão-somente aqueles que não foram ainda definidos no schema. Em um cromossomo com três genes teríamos uma figura tridimensional como mostra a

Figura 4.8.

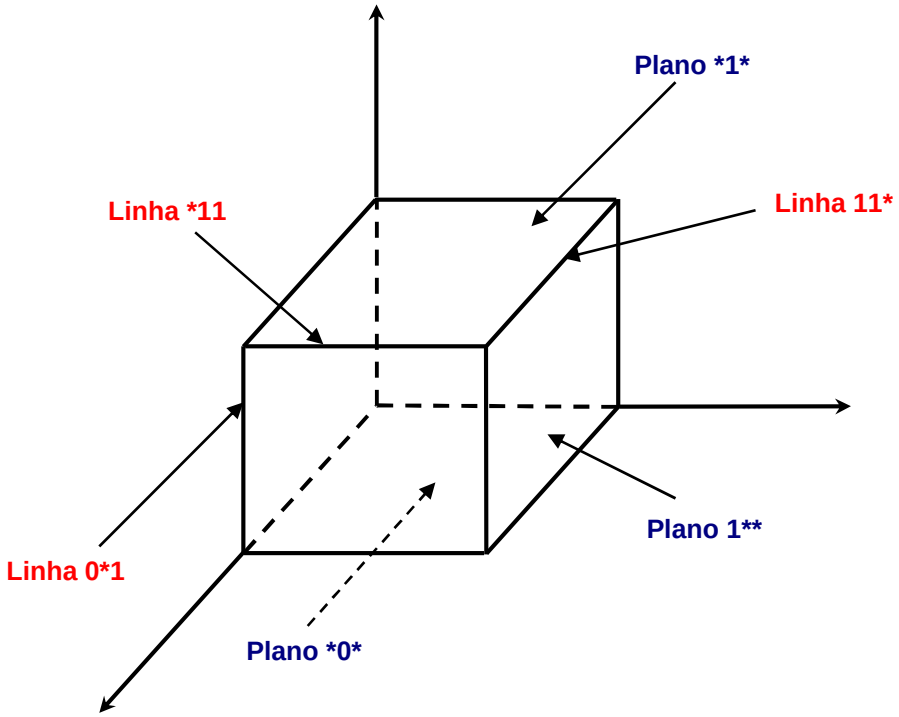

Figura 4.8 - Schemata como hiperplano em um espaço tridimensional

FONTE: GOLDBERG, 1989.

4.4 FUNDAMENTO

De uma maneira bastante geral, Mitchell (1996) explica o objetivo principal de AGs da seguinte forma: “Algoritmos genéticos trabalham descobrindo, enfatizando e recombinando bons ‘blocos de construção’ de soluções na mais alta forma de paralelismo”. Segundo Goldberg (1989), embora pareça muito simples, é justamente a simplicidade de operação e o poder de efetividade que representam a principal atração para o uso dos algoritmos genéticos.

4.5 ADEQUAÇÃO DO USO DE ALGORITMO GENÉTICO PARA O PROBLEMA

Para que se possa fazer uso de algoritmos genéticos é preciso analisar a possibilidade de adequação do problema à abordagem evolucionária. Basicamente, deve-se tentar modelar o AG considerando os passos descritos a seguir.

- *A representação das soluções*: é possível codificar candidatos à solução na forma de um cromossomo? Os alelos serão binários ou não? Qual o intervalo de valores possível?
- *O método de seleção*: qual o método de seleção que mais eficazmente é capaz de manter e melhorar o conteúdo genético[9] desejado?
- *Operadores genéticos*: que operadores genéticos realmente são necessários na evolução da população de soluções?
- *Definição dos parâmetros*: qual o tamanho da população, as condições de parada e a probabilidade de atuação dos operadores genéticos?
- *A função de fitness*: existe uma lógica em função da qual é possível orientar a busca do melhor indivíduo?

A seguir, cada um desses aspectos é descrito, permitindo conhecer o quão complexas são as opções de modelagem de um AG para que ele trabalhe um determinado problema com o maior desempenho, robustez e confiabilidade.

4.6 CODIFICAÇÃO E REPRESENTAÇÃO DO CROMOSSOMO

A forma de codificação do cromossomo é uma questão bastante discutida na literatura. Há autores a favor da codificação binária e há os que são contra ela, além de haver a codificação em forma de árvore (KOZA, 1992), a qual ainda não é tão popular.

Mitchell (1996) expôs essa questão, e, segundo a autora, trata-se de um fator central (se não “o fator central”) para o sucesso de um AG. A autora apresenta exemplos na literatura do uso de alfabetos com muitos caracteres e de números reais: gramáticas de geração de grafos, conjuntos de condições com valores reais, representação com números reais de pesos de Redes Neurais, representação com números reais para ângulos torcidos em proteínas, etc. Goldberg (1989) demonstra e testa apenas *schemata* que utilizam a codificação de bits.

O argumento de Holland (1975) implica que um alfabeto formado por muitos caracteres deveria apresentar uma pior performance. Embora a codificação binária seja a mais utilizada

[9] Conteúdo Genético: Conjunto de genes e alelos de um indivíduo, genoma.

(por razões históricas, pelo fato de métodos originais serem utilizados para alfabeto binário, etc.), ela também é uma forma não natural de representação (MITCHELL, 1996). Há comparações empíricas entre os dois tipos de codificação que mostraram uma melhor performance para o uso de valores reais, conforme apontam Janikow (1991) e Wright (1991).

Davis (1991), conhecido por aplicar AGs em situações do mundo real, defende fortemente que a codificação mais apropriada é aquela que melhor representa o problema que se pretende solucionar. Mesmo sabendo que geralmente os algoritmos genéticos trabalham com somente um tipo de codificação, o autor aconselha que o alfabeto seja escolhido primeiramente e só depois seja selecionado qual o melhor AG capaz de processar tal codificação.

Quanto à representação, a maior parte das aplicações de AGs utiliza indivíduos haplóides e que contêm apenas um cromossomo (MITCHELL, 1996). Apesar de cada gene poder ter vários alelos, também é mais comum utilizar somente genes com um alelo, isto é, apenas um valor ao mesmo tempo para cada gene.

4.7 MÉTODOS DE SELEÇÃO PARA REPRODUÇÃO

Também chamado de “operador de reprodução”, o método de seleção é o que irá decidir quais indivíduos deverão passar seu código genético para a próxima geração e em que proporção eles reproduzirão novos descendentes (GOLDBERG, 1989). O método de seleção escolhido é uma das chaves para a robustez da aplicação do algoritmo, na medida em que está diretamente relacionado à qualidade e à rapidez com que a população evolui em direção à solução desejada. A seguir, descrevem-se alguns dos métodos mais conhecidos.

- Roleta: é o método mais comum (MITCHELL, 1996) e trata-se de dar a cada indivíduo uma fatia de um círculo, a roleta, em que o tamanho da fatia representa o *fitness* no indivíduo; a roleta gira tantas vezes quanto o número da população; o indivíduo escolhido na roleta é selecionado para fazer parte da próxima geração.
- Escalonamento Sigma (ou Corte Sigma, segundo Goldberg (1989)): o algoritmo baseia-se na média e no desvio-padrão do *fitness* da população para dar chance a

cada indivíduo de ser selecionado ou não; a vantagem é poder manter a variedade da população no início do processo, quando o desvio-padrão do *fitness* individual ainda é grande em relação à população, tanto para indivíduos com pequeno *fitness* quanto para os de alto *fitness*.

- Elitismo: foi introduzido por De Jong (1975). Existem hoje muitas modificações para a implementação de elitismo, mas a idéia principal do método é garantir que os melhores indivíduos façam parte da próxima geração; assim, uma certa proporção de indivíduos com mais alto *fitness* é sempre mantida para constituir a próxima primavera (*offspring*).
- Seleção por ranking: proposta por Baker (1985), efetua a seleção de indivíduos através de uma escala construída a partir de seu *fitness*, ou seja, em vez de usar o valor absoluto do *fitness*, esse método utiliza um valor seqüencial, eliminando problemas com alta variância de *fitness* dentro da população, visto que não considera o quão longe está o valor de *fitness* de um e de outro indivíduo.
- Seleção por torneio: aqui, as chances de o melhor indivíduo ser escolhido são parametrizadas (valor entre 0 e 1); dois (ou mais) indivíduos são selecionados da população ao acaso; um número aleatório é sorteado e, se for menor que o valor parametrizado, o indivíduo com maior *fitness* no grupo fará parte da próxima geração, do contrário, o menos apto é selecionado; eles retornam para a população e podem ser selecionados novamente, até que toda a *offspring* esteja completa.
- *Steady-State*: funciona de modo quase inverso ao elitismo; neste método apenas uma pequena parte da população (formada pelos menos aptos) é substituída na geração seguinte; a substituição é feita por indivíduos criados a partir de mutação e *crossover* daqueles com mais alto *fitness*.

Ainda há diversos outros métodos, os quais resultam inclusive de combinações e variações dos aqui citados. O mais importante, no entanto, é saber qual deles melhor se adapta ao problema a ser solucionado. Ao comparar os métodos de seleção, Mitchell (1996) afirma que cada cálculo extra para a geração de uma *offspring* representa significativo consumo de tempo e processamento. Sabendo que até mesmo esse aspecto influencia na performance do AG, é preciso também considerar tal questão no momento de definir como o algoritmo genético deverá reproduzir.

4.8 OPERADORES GENÉTICOS

Os operadores genéticos representam o conjunto de fenômenos que, atuando paralelamente, resultam na evolução da população atual. Eles trabalham sobre o conteúdo genético dos indivíduos da população para a geração da próxima primavera, sempre com o objetivo prioritário de produzir indivíduos melhores. No entanto, segundo Mitchell (1996), o importante é o correto equilíbrio entre os operadores, o qual por sua vez depende da função objetivo e da codificação. A seguir são descritos os operadores mais comumente encontrados.

4.8.1 *Crossover*

A operação de *crossover* ou cruzamento é tida como a principal diferença entre o AG e as outras técnicas (MITCHELL, 1996). Trata-se da troca de segmentos de código genético (alelos) entre dois indivíduos com o mesmo genótipo (mesma espécie). O objetivo do *crossover* é recombinar características de indivíduos com alto *fitness* para gerar indivíduos mais aptos na próxima população. Esse operador é executado depois de feita a seleção de quais cromossomos terão seu conteúdo genético propagado na nova *spring*.

São feitos pontos de corte no cromossomo para aplicar o *crossover*. Os tipos diferentes de ponto de corte a serem adotados dependem, de maneira complexa, da função objetivo, da codificação e de outros detalhes do algoritmo utilizado (MITCHELL, 1996). A seguir são descritos alguns deles.

a) O ponto de cruzamento (*single-point crossover*) usa geralmente a seleção randômica para escolher em que altura o cromossomo sofrerá o corte. Desse ponto, o material cromossômico é trocado com outro indivíduo na geração de um novo cromossomo. Um exemplo[10] pode ser visto na Figura 4.9.

[10] Todas as figuras com exemplos de *crossover* apresentados utilizarão cromossomos com codificação binária para maior simplificação.

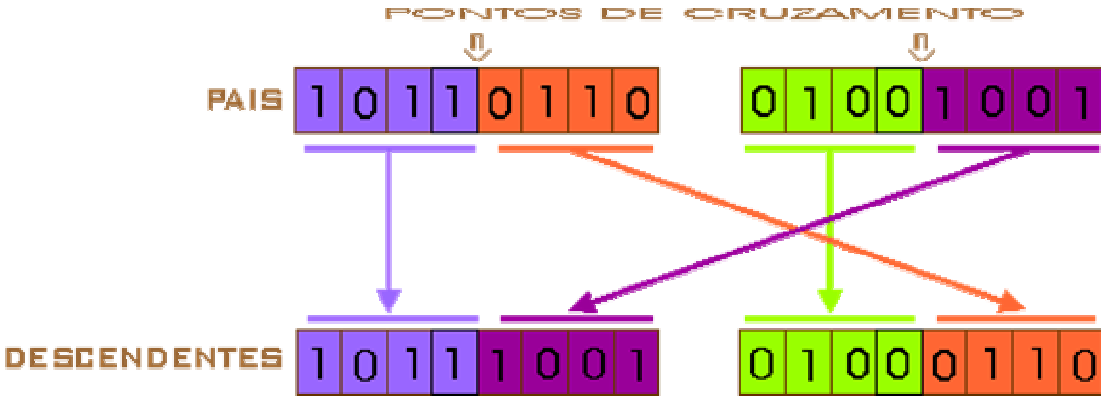

Figura 4.9 - *Crossover* de um ponto de cruzamento

FONTE: YEPES, 2004.

b) Quando o *crossover* utiliza a forma de dois pontos de cruzamento (*two-point crossover*), demonstrado na Figura 4.10, a troca de genes ocorre a partir dos dois pontos selecionados para corte, em que um dos cromossomos-pai contribui com dois trechos de sua *string* para um dos indivíduos descendentes e para outro descendente com somente um trecho.

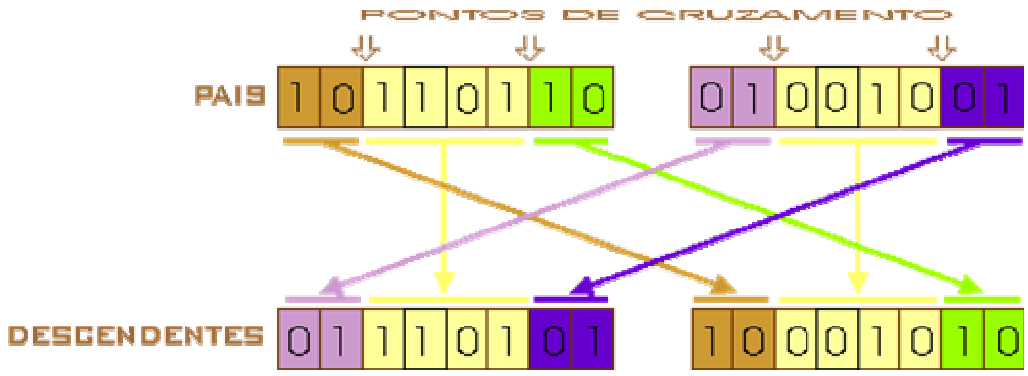

Figura 4.10 - *Crossover* de dois pontos de cruzamento

FONTE: YEPES, 2004.

c) O cruzamento uniforme é bastante diferente dos demais, utiliza-se de uma máscara para decidir quais os genes que serão trocados, sem usar a aleatoriedade. Não há número fixo dos pontos em que o cromossomo será cortado para troca, mas se costuma tomar como base o comprimento do indivíduo para se decidir. O exemplo é mostrado na Figura 4.11.

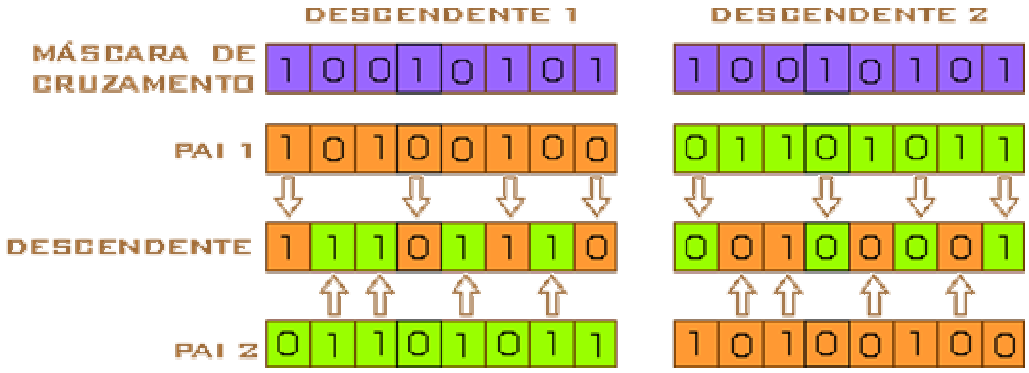

Figura 4.11 - Cruzamento uniforme

FONTE: YEPES, 2004.

4.8.2 Mutação

Juntamente com o operador de *crossover*, a mutação (ou inversão) é responsável pela diversidade, porém, mais especificamente, pela variedade e pela inovação do conjunto de cromossomos (MITCHELL, 1996). Esse operador produz valores aleatórios para os genes, podendo introduzir conteúdo genético inédito na próxima *spring*, isto é, fora do espaço de busca.

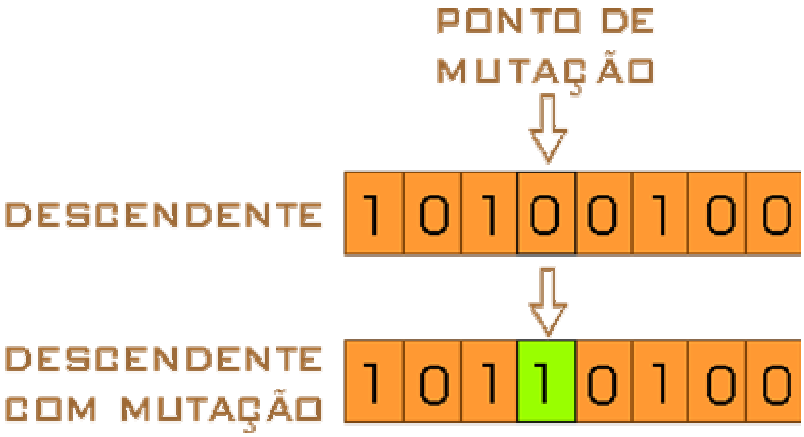

Figura 4.12 - Operador de Mutação

FONTE: YEPES, 2004.

Ao fazer uso de codificação não binária para o cromossomo, cada gene passa a ter um intervalo de possíveis valores, por exemplo, a característica “idade” somente pode conter valores entre 0 e 150. Nesse caso, é comum estruturar um mecanismo (no algoritmo ou no banco de dados) para que o operador de mutação possa consultar quais valores deve selecionar a partir de um conjunto de possíveis alelos, criando assim um domínio pré-

estipulado. Um exemplo simples é mostrado na Figura 4.12, na qual o cromossomo pertencente à nova *spring* sofre ação do operador de mutação.

4.8.3 Outros

Existem ainda muitos outros operadores genéticos, como, por exemplo, o operador *crowding*, introduzido por De Jong (1975), ou o *fitness sharing*, estudado por Goldberg e Richardson (1987), ou ainda o *mating tags*, de Holland (1975) e Booker (1985). Todos esses operadores com características específicas variadas procuram melhorar a diversidade da população e equilibrar a rapidez de sua convergência. Porém, o programador poderá implementar combinações desses operadores ou criar novos, conforme a necessidade e o ambiente de aplicação.

4.8.4 Parametrização

Quanto à parametrização dos operadores, o usuário poderá definir quais deseja manter variáveis ou deixar fixos no código do algoritmo, conforme o problema. Selecionar os valores para os parâmetros é a quarta tarefa na implementação de um AG (MITCHELL, 1996). Em geral, informam-se parâmetros para (SALVADOR, 2000):

- tamanho da população: é uma característica que afeta o desempenho e a eficiência do algoritmo, determinando a cobertura do espaço de busca, a rapidez da convergência e a necessidade de recursos computacionais;
- taxa de mutação: define a capacidade de inovação das soluções, isto é, possibilita que os indivíduos atinjam qualquer ponto no espaço de busca, porém, altos valores de mutação tornam a busca aleatória;
- taxa de *crossover*: determina a velocidade com que novas estruturas surgem na população – baixas taxas causam lenta variação dos indivíduos, enquanto um alto valor para esse parâmetro pode fazer com que estruturas com alta aptidão sejam perdidas; e
- intervalo de geração: trata-se da porcentagem da população que será substituída por novos indivíduos a cada *spring*.

4.9 FUNÇÃO OBJETIVO

A função de *fitness,* como a função objetivo como é chamada pelos biólogos, representa uma medida que desejamos maximizar. Trata-se da versão artificial para a seleção natural de Darwin em que o *fitness* de um indivíduo representa sua habilidade de sobreviver a predadores, doenças e outros obstáculos. No paralelo artificial, a função objetivo é que decidirá quais as chances de o indivíduo “viver” ou “morrer” (GOLDBERG, 1989).

4.10 FUNCIONAMENTO

Os mecanismos básicos de funcionamento de um AG são surpreendentemente simples, segundo Goldberg (1989). Após decidir quais serão a representação e a codificação do cromossomo, gera-se aleatoriamente uma população, de tamanho fixo ou variável, de potenciais soluções para um problema qualquer. Sobre essa população aplicam-se operadores genéticos parametrizados, dos quais se espera que causem a evolução dessa população na direção que se deseja. Depois que os operadores genéticos agem sobre os indivíduos, ocorre a reprodução de acordo com o método de seleção escolhido.

4.11 DIFERENÇAS ENTRE ALGORITMOS GENÉTICOS DOS MÉTODOS TRADICIONAIS

Segundo Goldberg (1989), os algoritmos genéticos – diferentes de outros métodos de busca e otimização – possuem quatro aspectos principais:

1) operam com a codificação do conjunto de parâmetros em vez de trabalharem diretamente com parâmetros;
2) buscam uma população de pontos e não um único ponto no espaço de busca;
3) utilizam-se da informação gerada pela função *payoff* (função objetivo) e não de derivadas ou conhecimento auxiliar; e
4) fazem uso de regras de transição probabilística em vez de regras determinísticas.

Levando em conta o escopo de aplicações de técnicas de extração de conhecimento, Romão (1999) destaca que a principal motivação para o uso de AGs na extração de regras de previsão reside no fato de que algoritmos genéticos são capazes de considerar a interação entre atributos no processo de busca, característica, segundo ele, crucial para o sucesso de tais técnicas.

Além disso, já foram provadas teórica e empiricamente a robustez e a eficiência de algoritmos genéticos na busca de soluções ótimas em espaços complexos de problemas com muitos atributos (XIONG; LITZ, 1999).

Cabe aqui dizer que na escolha das ferramentas e técnicas para aplicar mineração de dados vários aspectos do problema devem ser analisados (recursos disponíveis, necessidades de negócio, etc.) (INMON et al., 2001), mas:

> Não há um método de Mineração de Dados ‘universal’ e a escolha de um algoritmo particular para uma aplicação particular é de certa forma uma arte. (FAYYAD et al. , 1996b, p. 86).

4.12 MÉTODOS DE BUSCA

Quando se têm um problema e um espaço constituído de diversas soluções possíveis, resolver esse problema é apenas uma questão de encontrar a melhor solução ou uma solução ótima entre as existentes no escopo disponível. Para isso, existem diversos métodos de busca. O desafio é selecionar aquele que melhor se adapta ao contexto, apresentando maior nível de desempenho e eficácia.

Goldberg (1989) analisa – sem fazer testes formais – os três tipos de método de busca identificados na literatura, descrevendo suas vantagens e desvantagens, como apresentado a seguir.

a) Métodos baseados em cálculos

Dividem-se em duas classes principais: diretos e indiretos. Os indiretos procuram por extremos locais, resolvendo um conjunto de equações lineares resultantes do gradiente da função objetivo, quando é igual a zero. Os métodos diretos procuram por ótimos locais e

escalam o gradiente local utilizando a função dada (técnica de *hill-climbing*). Este método tem a desvantagem de encontrar o máximo local e acabar perdendo o máximo global; além disso, como o próprio nome diz, é um método que requer a existência de derivadas (valores bem definidos de subida do gradiente), mas os dados no mundo real são muitas vezes descontínuos, ausentes e multimodais. Essas características acabam por restringir o domínio de uso do método.

b) Métodos enumerativos

Dado um espaço de busca finito ou discretizado infinito, essa técnica utiliza uma função objetivo em cada ponto existente, um de cada vez. Embora simples, o problema evidente deste método é a eficiência, visto que muitos problemas práticos possuem um espaço de busca grande demais para que se possa analisar todos os seus pontos – o que Bellman (1961) chamou de “maldição da dimensionalidade”.

c) Métodos randômicos

Antes de tudo é preciso explicar que há diferença entre métodos randômicos e técnicas randomizadas; estas últimas utilizam-se da opção aleatória para guiar uma busca de exploração pela codificação de um espaço parametrizado. No tocante aos métodos aleatórios, apesar de populares, ao longo do uso podem ser tão ineficientes quanto os enumerativos, já que utilizam o acaso para fazer buscas, não tendo qualquer direcionamento dessa procura.

Goldberg (1989) ressalta que os AGs, por explorarem similaridades de várias formas, tornam-se bem menos restringidos pelas limitações que afetam outros métodos, tais como continuidade, existência da derivada, busca ao acaso, etc.

A seleção natural tem como vantagens a solidez e o paralelismo herdado (GUIMARÃES, 2003), mas possui desvantagens quanto à geração de indivíduos, como, por exemplo, um classificador genético, que necessitaria de um número muito maior de exemplos de treinamento para alcançar resultados semelhantes aos alcançados por árvores de decisão (LUCAS, 2002).

4.13 APLICAÇÕES DE ALGORITMOS GENÉTICOS

A seguir são relacionadas as aplicações mais comuns de AGs seguidas dos campos, de áreas e dos ambientes onde são praticadas.

- *Otimização*: otimização numérica, design de circuitos e escalonamento.
- *Programação automática*: utilizada para desenvolver programas para tarefas específicas e outras estruturas computacionais, tais como autômatos celulares.
- *Aprendizagem computacional*: classificação e previsão (meteorologia), aprendizagem dos pesos de redes neurais, regras de sistemas de classificação/produção e robótica.
- *Economia*: estratégias de definição de preços.
- *Sistemas sociais*: utilizados para estudar a evolução do comportamento social, a evolução da cooperação e comunicação em sistemas multiagentes.
- *Biologia*: estudo do sistema imunológico e da relação entre a aprendizagem individual e a evolução das espécies.

Para uma relação detalhada e mais extensa sobre o uso de AGs em diferentes áreas de aplicação, ver Goldberg (1989, p. 126-129).

4.14 CONSIDERAÇÕES FINAIS

Neste capítulo efetuou-se o levantamento bibliográfico sobre Algoritmos Genéticos quanto a conceitos evolucionários, características gerais, funcionamento e aplicações, discutindo-se brevemente as motivações para o seu uso e para possíveis configurações.

A seguir são descritos os experimentos de mineração de dados, objetivando a busca de regras de classificação em amostras de dados de uma rede de baixa tensão, utilizando-se algoritmo genético. As questões levantadas aqui a respeito da modelagem e da escolha do algoritmo são colocadas em prática, permitindo testar a efetividade da abordagem evolucionária no escopo do problema, as possíveis variações quanto aos parâmetros e à performance, além de conhecer melhor o ambiente de exploração.

5 TESTE COM A ABORDAGEM EVOLUCIONÁRIA

Até este ponto foram apresentados neste trabalho as revisões bibliográficas – para dar embasamento aos estudos desenvolvidos – e o ambiente do problema. Neste capítulo, são definidos em detalhes os processos necessários para que o trabalho fosse desenvolvido, desde a obtenção dos dados das bases operativas e o processamento das amostras de dados até a alteração do algoritmo genético utilizado e a configuração de seus parâmetros para a realização de testes.

Inicialmente, um algoritmo genético simples foi implementado. O objetivo foi testar a capacidade da base de dados de ser preparada para o uso de AGs, assim como avaliar a abordagem de Inteligência Artificial quanto à geração de regras de previsão. Tendo atingido sucesso na preparação dos dados, conforme são requeridos para a apropriada extração do conhecimento (seguindo os passos do processo KDD, descritos na página 50, Figura 3.5) e para encontrar regras de classificação válidas, o trabalho estende-se para selecionar e testar um algoritmo genético mais complexo, cujas características possam se adaptar melhor às necessidades gerenciais em constante mudança.

Além de permitir conhecer a adequação da técnica para o cenário do problema, o teste inicial serviu para aprofundar o conhecimento sobre as regras de negócio e os aspectos computacionais envolvidos, bem como para indicar as potenciais fontes de informação, as relações entre elas e sua relevância dentro do contexto do trabalho. As atividades desenvolvidas durante o teste serão parcialmente aproveitadas para a aplicação do algoritmo mais complexo.

Este capítulo inicialmente apresenta o cenário do problema e do ambiente no qual o trabalho foi desenvolvido. Em seguida um AG simples é testado no contexto de baixa tensão; sendo descritos brevemente a modelagem dos dados necessários para sua aplicação e os resultados alcançados neste estudo de caso. Por fim, descreve-se o algoritmo genético complexo que foi selecionado para efetuar a mineração deste estudo; são relatadas as requeridas adaptações e modificações quanto ao código do AG, a preparação de amostras de

dados para seu uso e os experimentos realizados na tentativa de encontrar regras de classificação relevantes, úteis e de qualidade.

5.1 CENÁRIO DE APLICAÇÃO

O ambiente de informações de uma rede de distribuição elétrica constitui uma vasta área para exploração e desenvolvimento de tecnologias voltadas para o setor estratégico. As características complexas do comportamento dos circuitos elétricos, em conjunto com a diversidade de aspectos externos que interferem sobre na operação e no controle de seus componentes, demandam inerentemente conhecimento que auxilie na otimização de processos, na eficácia das soluções implementadas, na orientação quanto ao direcionamento de recursos, entre outras tarefas de tomada de decisão.

Ao inserir-se em um projeto já existente na CELESC, este trabalho de mineração de dados se beneficia do fato de ter à disposição um *data warehouse* com *data marts* sobre as redes de distribuição de energia. Também conta com a experiência de especialistas já envolvidos com o DW e que possuem interesse na busca por conhecimento aplicável em suas áreas. Em contato com esses especialistas se descobre rapidamente a infinidade de problemas de gerenciamento da área elétrica e para os quais ainda há solução prática. Muitos desses casos podem ser, no mínimo, auxiliados por técnicas e ferramentas de software.

É com essa motivação que este estudo faz uso do ambiente de informações das redes de distribuição de energia, integrando a experiência de engenheiros conhecedores do domínio do problema e a mineração de dados para alcançar soluções que contribuam relevantemente para o trabalho que eles desenvolvem e, conseqüentemente, para aqueles que se utilizam dos serviços prestados pela companhia elétrica.

Para os experimentos realizados neste estudo, selecionou-se o problema das interrupções elétricas no fornecimento de energia. A principal razão para essa escolha são as implicações financeiras relacionadas às resoluções da ANEEL que estabelecem metas (item 2.4.2) e impõem multas (item 2.4.4) sobre as concessionárias de eletricidade por violação dessas metas. Desse modo, ajudar a prevenir interrupções de energia significa não apenas

gerar qualidade de fornecimento para os consumidores, mas também diminuir as perdas de receita e reduzir os custos com o pagamento de penalidades por descontinuidade da distribuição.

O uso de um algoritmo genético para a extração de regras de classificação nesse ambiente de aplicação é motivado principalmente pela idéia de se encontrar conhecimento relevante e não trivial, dando prioridade para que os resultados obtidos sejam compreensíveis a quem faz uso dele, pois este também está entre os aspectos principais na definição de *data mining* (FRAWLEY et al., 1992). O método utilizado objetiva fornecer apenas regras de classificação conhecidamente interessantes ao usuário e, para isso, ele é guiado em uma busca paralela e evolutiva através das hipóteses levantadas por especialistas sobre o domínio de informação. Dessa maneira e por meio de uma complexa avaliação da qualidade da regra, o método não sobrecarrega o usuário com classificações que fogem de seu escopo de análise, nem tampouco mascara ou inviabiliza a localização das regras de verdadeiro interesse para ele.

Tendo em vista a reutilização da ferramenta para a aplicação em novos problemas de baixa tensão, buscou-se o máximo possível de autonomia na solução de software. Nos testes feitos se procurou também simular a situação mais próxima da realidade, estando dependente do conhecimento do usuário para deixá-lo direcionar os objetivos da mineração e mostrar como os resultados seriam aplicados. A validação pelos usuários das regras de classificação geradas neste estudo é a continuação dessa abordagem, mas também visa principalmente incentivá-los a fazer uso da tecnologia pesquisada.

5.2 TESTE COM A ABORDAGEM EVOLUCIONÁRIA

Para validar a abordagem evolucionária ao ambiente de dados das redes de baixa tensão, foi aplicado um algoritmo genético simples a pequenas amostras de dados de circuitos elétricos. O objetivo deste estudo teve, entre suas prioridades a simplicidade na implementação, pois se pretendia basicamente testar e comprovar se o uso de algoritmos genéticos podia alcançar ótimos resultados na geração de regras sobre dados de baixa tensão.

5.2.1 O algoritmo genético

O algoritmo genético utilizado para o teste é uma adaptação do algoritmo descrito por Goldberg (1989), cujo código em linguagem Pascal é apresentado na mesma referência. As adaptações necessárias foram feitas utilizando-se a linguagem de programação Delphi 7.0. O programa principal apresentando a arquitetura hierárquica e de controle do software está descrito a seguir.

```
     begin {programa principal}
gen := 0;
initialize;
statistics (popsize, max, avg, min, sumfitness, oldpop);
repeat until (gen >= maxgen)
 begin
 gen := gen + 1;
 generation;
 statistics (popsize, max, avg, min, sumfitness, newpop);
 oldpop := newpop;
 end;
end.
```

Onde:

- gen: geração atual
- maxgen: número predefinido de gerações
- popsize: tamanho predefinido da população
- max: *fitness* máximo da população
- avg: média do *fitness* da população
- min: *fitness* mínimo da população
- sum*fitness*: somatório do *fitness*
- oldpop: antiga geração da população
- newpop: nova geração da população

O primeiro procedimento – *initialize* – executado no algoritmo é a inicialização da população aleatória e sem repetição de indivíduos segundo o tamanho que lhe foi determinado (*popsize*). Em seguida, a função *statistics*, uma função de avaliação, analisa a população inicial extraída calculando os parâmetros que irão medir a qualidade dos indivíduos na escolha das próximas gerações. Tendo uma população inicial, o algoritmo inicia o processo repetitivo para otimizar as possíveis soluções no ambiente do problema. Isso é feito produzindo-se novas gerações da população de acordo com o valor do *fitness* alcançado por elas.

É no processo de geração, efetuado pelo procedimento *generation*, que ocorre a aplicação do *crossover* e da mutação, de acordo com as probabilidades informadas pelo usuário. Esse ciclo termina quando o número de gerações é alcançado – preferiu-se deixar aos testes e não ao próprio algoritmo a análise quanto à significância da variação obtida na

população ou sobre a eficácia dos operadores genéticos de acordo com a probabilidade determinada a cada um. Com esse método, os melhores indivíduos deram origem à nova população na medida de dois cromossomos antigos para dois novos cromossomos, modificados geneticamente na tentativa de atingir um maior espaço no escopo de soluções possíveis.

5.2.2 Definição dos aspectos genéticos

Quanto ao indivíduo, sua representação é haplóide (seção 4.6). Nesse caso, cada cromossomo representa um circuito de baixa tensão, e cada característica do circuito (índice de carregamento, quantidade de unidades consumidoras por classe, fator de potência, etc.) é um gene do cromossomo. A codificação não é binária, para tornar possível aproveitar melhor os intervalos de valores de cada atributo do circuito.

A função *payoff* implementada para esse problema é bastante simples e é diferente daquela utilizada no algoritmo original, pois precisou adequar-se ao contexto do problema aqui descrito. Baseia-se em executar uma consulta SQL ao banco de dados, tendo como restrição (em sua cláusula "where") o conseqüente da regra. A partir desse conjunto de dados, cada aspecto do circuito no banco de dados é comparado com o correspondente aspecto no indivíduo da geração. Para cada gene com valor igual ao valor do respectivo atributo é acrescentado um ponto ao seu *fitness*. Se o indivíduo na sua totalidade for encontrado no conjunto de registros trazido do banco de dados, vinte pontos são acrescentados, premiando assim indivíduos cujos valores alcancem dados reais.

Quanto ao tamanho da população, decidiu-se utilizar cerca de um terço dos registros do conjunto de dados para compor a população processada pelo algoritmo. Feito desse modo, 1.000 registros são utilizados como população, enriquecendo em muito a diversidade possível na evolução dos indivíduos. Os benefícios alcançados por essa abordagem refletem-se no *fitness* máximo obtido quando do uso de apenas 100 indivíduos na população em comparação com a base completa (3.072 registros).

Sobre o valor parametrizado para os operadores, escolheram-se, após vários testes, os valores de 60% para *crossover* e 8% para mutação. À medida que esses valores foram

elevados, percebeu-se que os indivíduos com mais alto *fitness* eram encontrados menos vezes pelo algoritmo. Esse aspecto, de modo inverso, também se refletiu pelo número de gerações selecionado para o processo, ou seja, o algoritmo convergia para aquele cromossomo, passando a gerar populações constituídas do mesmo material genético.

Para o parâmetro Tamanho da população, considerou-se o discutido na seção 4.8, buscando-se alcançar um valor para o tamanho populacional que permitisse boa cobertura do espaço de busca e um ótimo uso dos recursos computacionais, bem como que não gerasse problemas de convergência prematura.

5.2.3 Preparação dos dados

Como é usual em processos de *Data Mining*, neste trabalho também foi preciso executar a preparação dos dados, de forma a evitar *outliers*, dados ausentes e perdidos, etc. (item 3.3.5). Para alcançar o melhor desempenho do algoritmo genético utilizado, foi necessário tratar os dados visando ao uso desta técnica em particular. Isso exigiu trabalho com dados discretizados, suporte a dados ausentes, entre outros fatores.

a) Seleção

Esta tarefa foi necessária para definir, num primeiro estágio, quais características dos circuitos de baixa tensão eram pertinentes à análise pretendida e se possuíam alguma possível contribuição para a obtenção das regras. Os dados foram extraídos do *Data Warehouse* implantado na CELESC, num total de 5.210 registros. Dos cento e vinte atributos, apenas trinta foram selecionados pelos especialistas em redes de distribuição de energia para análise como sendo interessantes e tendo envolvimento com as classes desejadas pela mineração. A seleção dos atributos pelos especialistas foi feita de modo empírico, segundo conhecimento prévio do ambiente do problema; análises de correlação ajudaram a refinar essa seleção.

b) Pré-Processamento

Apenas 3.072 indivíduos adequados foram encontrados a partir da amostra original. Muitos registros não possuíam valores quanto aos atributos selecionados. Para processá-los, foi criada uma estrutura de banco de dados, equivalente a um *Exploration Warehouse*, descrito na seção 3.2.5. O objetivo dessa estrutura era fazer pesadas análises estatísticas sobre

os registros, descobrindo possíveis padrões e relacionamentos entre os dados, sem interferir no processamento existente no *data warehouse* corporativo.

c) Transformação

Após se certificar de que o conjunto era composto apenas de valores válidos, foi necessário prepará-los para oferecerem o máximo de significância e robustez à execução do algoritmo genético e à solução por ele gerada. Entre as transformações feitas, trabalhou-se sobre o atributo correspondente ao desequilíbrio do circuito, estabeleceu-se a quantidade de trechos em faixas predeterminadas de queda de tensão e classificou-se o circuito quanto à porcentagem de queda de tensão presente nele. Também foi encontrado o número mais adequado de classes (doze faixas), segundo a fórmula de Sturges, sugerida em Pacitti et al. (1977): 1 + 3.3 log10N, onde N é o tamanho da amostra.

d) Seleção dos atributos para as regras

A definição dos atributos que comporiam o cromossomo, ou seja, as características que comprovadamente contribuiriam para uma regra útil, foi feita através de uma análise estatística descritiva, examinando-se a correlação de todos os atributos com o atributo conseqüente da regra. Aqueles campos que atingissem mais de 50% de correlação com as variáveis independentes eram selecionados para formar o cromossomo que serviria para a regra. O procedimento foi repetido para todas as classes das quais se desejava extrair regras. A programação no software também foi feita conforme esses campos selecionados. Desse modo, o algoritmo processa os resultados baseando-se apenas nos atributos comprovadamente pertinentes à regra selecionada.

e) Aplicação do algoritmo

O algoritmo foi aplicado repetidamente, testando-se diversas combinações de valores para os parâmetros e modificações no tocante à função objetivo. Por se tratar de um software desenvolvido para protótipo e não para real uso, a performance alcançada foi considerada razoável, principalmente tendo em vista a configuração mediana da máquina utilizada para a execução do algoritmo e o tamanho significativo da população parametrizada.

5.2.4 Resultados alcançados

Obtiveram-se três regras de classificação com alto *fitness* dentro do conjunto de dados. Essas regras foram apresentadas a especialistas da CELESC que as confirmou como válidas.

Esses resultados não apenas conferiram a capacidade da abordagem evolucionária na tarefa de encontrar regras de classificação válidas sobre dados de baixa tensão, mas também permitiram no contexto do problema: a análise de desempenho de um AG sobre atributos característicos; o estudo da adequação quanto à representação e codificação cromossômica; os tipos e a validade da amostragem, o nicho de interesse dos especialistas; a familiarização com os sistemas operativos e com determinados aspectos reais do ambiente; a exploração e o levantamento de hipóteses à medida que os dados iam sendo trabalhados, entre outros muitos detalhes.

Após esse teste bem-sucedido do uso de algoritmos genéticos para a mineração de dados no ambiente do problema deste estudo, pôde-se partir para a segunda etapa do trabalho: a busca de uma solução de software para a extração de regras de classificação em redes de baixa tensão. No capítulo a seguir descreve-se um algoritmo genético mais complexo, embutido em um sistema híbrido.

6 O ALGORITMO GENÉTICO DO SISTEMA AGD

O algoritmo genético selecionado foi desenvolvido por Wesley Romão em sua tese de doutorado (ROMÃO, 2002). O modelo implementado por ele como protótipo é um sistema denominado AGD, ou Algoritmo Genético para Descoberta de Regras Difusas. A sua pesquisa utilizou AG e Lógica Difusa para a tarefa de classificação, buscando a representação de regras através de indivíduos de um algoritmo genético. E é por envolver duas técnicas diferentes em sua estrutura que o AGD é considerado um sistema híbrido, nesse caso, híbrido-difuso.

São apresentadas a seguir as principais características do AGD de modo resumido, ressaltando os aspectos que afetarão direta e indiretamente o escopo deste trabalho. Mais detalhes sobre o algoritmo poderão ser encontrados na tese do autor.

6.1 ORGANIZAÇÃO DO SISTEMA AGD

Basicamente o algoritmo genético para a extração de regras difusas proposto por Romão (2002) é organizado para operar de acordo com a Figura 6.13, em que, a partir de uma *data warehouse*, atributos relevantes para a mineração de dados são integrados em um ambiente para aplicação do processo de extração de regras de classificação.

O AGD reúne as características de algoritmos genéticos – quanto à busca no espaço global de soluções e à consideração da interação existente entre os atributos – e conjuntos difusos – no que se refere à representação de valores contínuos através de termos linguísticos – para processar o conjunto de dados da mineração.

Guiado pelo cálculo da qualidade através da matriz de confusão (grau de pertinência) e do cálculo do grau de interesse (grau de similaridade), o AGD realiza a avaliação das regras obtidas medindo sua relevância (*interestingness* do resultado) (PIATETSKY-SHAPIRO; MATHEUS, 1994) para o usuário, de acordo com as impressões gerais (IGs) informadas plo mesmo. Por fim, o AGD objetiva obter conhecimento estratégico para a organização

combinando ao mesmo tempo validade, novidade, simplicidade (compreensibilidade) e utilidade desse conhecimento (ROMÃO, 2002).

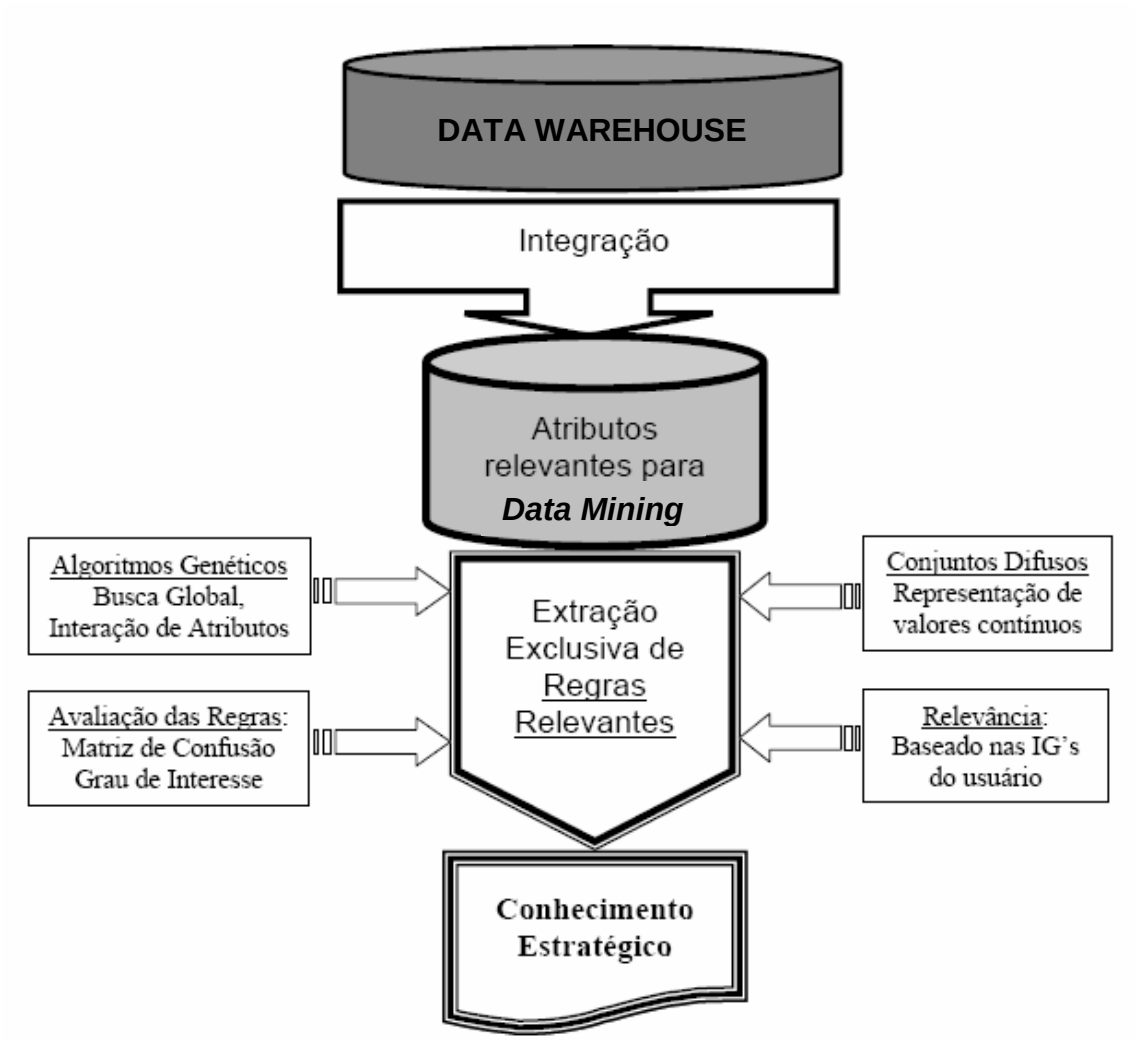

Figura 6.13 - Organização do Sistema AGD.

FONTE: adaptado de: ROMÃO, 2002.

6.2 CODIFICAÇÃO E REPRESENTAÇÃO DO CROMOSSOMO

A forma de representação utiliza-se da abordagem de Michigan (seção 4.1), isto é, cada indivíduo no algoritmo genético representa uma regra. A definição utilizada de regra não varia conforme o conceito geral (item 3.3.10). Na estrutura definida pelo autor, tanto antecedente como conseqüente da regra estão contidos no mesmo cromossomo.

Todos os indivíduos possuem tamanho igual e fixo, correspondente ao número de atributos da amostra de dados utilizados pela mineração. Dessa maneira, não é preciso modificar o cromossomo conforme a regra pretendida, visto que todos os indivíduos possuem o mesmo genótipo, embora seu fenótipo seja variável (Romão, 2002).

Os genes têm duplo alelo, um para o valor do gene e outro como *flag* para controle interno do AG. O valor do *flag* varia de 0 a 2, conforme as definições dispostas na Tabela 6.13. O nome do atributo é determinado pelo índice do gene, por isso não é necessário armazená-lo no genoma. O gene contendo o conseqüente da regra – ou atributo meta – é indicado por sua posição em um gene especial, com apenas um alelo, no final do cromossomo. Essa estrutura é apresentada na Figura 6.14.

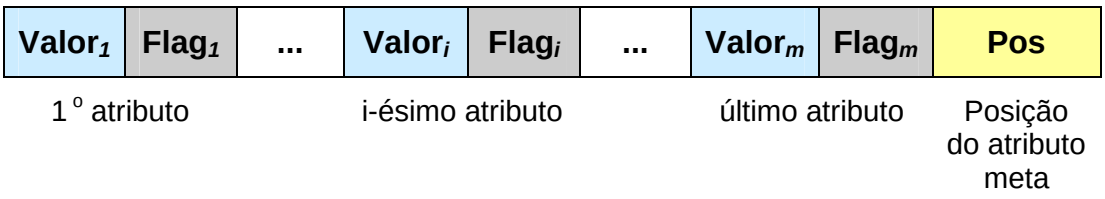

Figura 6.14 - Codificação do cromossomo

Onde:

i = o i-ésimo atributo da regra;

m = quantidade de atributos selecionados do banco de dados para a mineração;

$\boldsymbol{Valor_i}$ = valor do domínio do atributo i;

$\boldsymbol{Flag_i}$ = indica a ativação do gene no cromossomo, no antecendente ou no conseqüente.

Valores para o campo Flag	Significado dos valores para o sistema
Flag = 0	Atributo está desativado no antecedente e pode fazer parte do conseqüente.
Flag = 1	Atributo está ativo no antecedente.
Flag = 2	Atributo está desabilitado no cromossomo.

Tabela 6.3 - Significado dos valores de *flag* no gene

O tipo de codificação do gene é capaz de armazenar tanto atributos descritivos – ou categóricos – quanto atributos contínuos “fuzzificados”. Nesse sistema, não foi necessário cogitar a aplicação de um alfabeto binário para a codificação cromossômica (discutido na seção 4.6), o qual diminuiria a potencialidade dos dados quanto ao universo de soluções ao criar uma estrutura cujos alelos fossem compostos apenas de valores zero e um. Aqui, a lógica difusa embutida no sistema permitiu que valores contínuos fossem utilizados sem prejuízo para o desempenho do algoritmo, porque a técnica reduz o espaço de busca sem comprometer a variedade de soluções disponíveis.

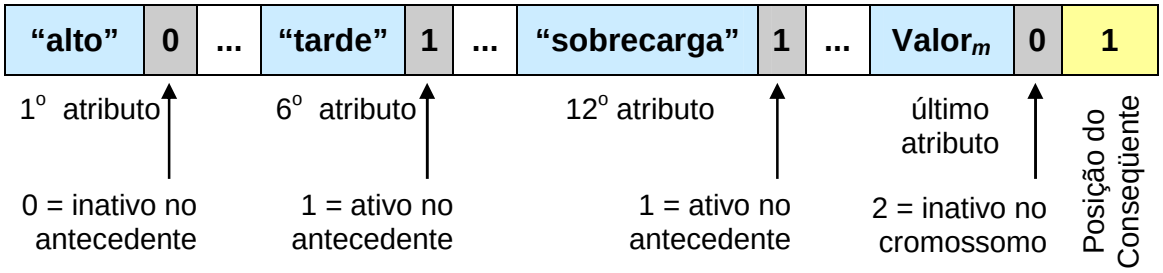

Figura 6.15 - Exemplo de codificação do cromossomo

A Figura 6.15 apresenta um exemplo de um cromossomo codificado com a seguinte regra: (Período = “tarde”), (Causa = “sobrecarga”) => IN_DEC = “alto”. O último gene indica que o gene de índice “1” é o gene que contém o atributo conseqüente da regra.

O operador utilizado para atributos categóricos e fuzzificados é o “=”, pois se trata de condições que fazem uso de valores *linguísticos* e podem ser expressas na forma “Atributo$_i$ = Valor$_i$”, como, por exemplo, “Tensão Nominal = Alta”.

6.3 SELEÇÃO DA POPULAÇÃO

O algoritmo genético do sistema AGD utiliza seleção por torneio (seção 4.7), coletando um número fixo de indivíduos (dois) para serem escolhidos conforme o maior *fitness* entre os eles. Em seguida, o mesmo processo busca outro indivíduo na população, garantindo que este segundo é diferente do primeiro já selecionado para reprodução. O elitismo é implementado passando-se os dois melhores indivíduos para a população seguinte sem que eles sofram modificações.

6.4 OPERADORES GENÉTICOS

Além dos operadores de cruzamento (*crossover*) e mutação, Romão (2002) desenvolveu outros dois operadores genéticos aplicáveis especificamente a esse tipo de estrutura de cromossomo – em que os genes podem ser ativados ou desativados no indivíduo. Os operadores comuns também sofreram algumas variações em relação à sua aplicação. Os operadores genéticos definidos para o AGD são descritos brevemente a seguir.

6.4.1 Crossover

Além do funcionamento comumente utilizado para realizar *crossover* (item 4.8.1), Romão (2002) aplicou uma pequena modificação devido à característica própria da codificação utilizada no cromossomo. Para evitar que atributos inativos fossem mais freqüentemente selecionados para cruzamento do que os ativos, o autor implementou um fator de probabilidade interno ao cromossomo. Esse fator permite a distribuição uniforme da ação do operador sobre os atributos ativos pertencentes ao antecedente da regra.

O valor parametrizado para o *crossover* foi definido empiricamente como 85%. O fator de probabilidade interna de cruzamento ficou em 50%. Segundo Romão (2002), esses valores alcançaram resultados preliminares satisfatórios, evitando convergência prematura e reduzindo o número de indivíduos repetidos na população.

6.4.2 Mutação

O operador de mutação (item 4.8.2) altera somente o valor de atributos ativos dentro da regra, de acordo com um índice de probabilidade parametrizado, ignorando atributos inativos ou desabilitados e mantendo a quantidade de condições ativas (fenótipo). Os valores mutados variam dentro do intervalo válido àquele atributo, conforme uma tabela de domínio. Também por determinação empírica, Romão (2002) estimou a probabilidade para ocorrer mutação em 2%.

Neste operador também foi desenvolvido um segundo fator de probabilidade: a probabilidade específica de mutação sobre um atributo ativo. Esse fator é calculado dinamicamente conforme o tamanho do domínio de cada atributo, permitindo assim que genes

com um domínio extenso sofram mutação mais freqüentemente. O cálculo baseia-se na seguinte relação:

$$ProbEspecifica_i = (TamDom_i - 1) * ProbMutação$$

Onde:

ProbEspecifica$_i$ = Probabilidade específica do atributo i;

TamDom$_i$ = Tamanho do domínio do atributo i;

ProbMutação = Probabilidade geral de ocorrer mutação.

6.4.3 Operadores de inserção e remoção de condições

Ambos os operadores têm a função de melhorar a compreensão das regras e aumentar a diversidade da população através da exploração do espaço de busca (Romão, 2002). Esses operadores atuam diretamente sobre o número de condições (atributos) da regra, isto é, sobre o antecedente da regra, sendo executados após a operação do *crossover* e da mutação. A probabilidade de ocorrerem é parametrizada do mesmo modo que para os demais operadores genéticos e foi definida pelo autor empiramente.

Porém, diferentemente dos demais operadores, em vez de trabalharem sobre o valor do gene, eles atuam sobre o alelo “Flag”, trocando seu conteúdo de 0 para 1, e vice-versa. Dessa maneira, eles ativam ou desativam genes no cromossomo. Seu funcionamento obedece ao parâmetro do sistema, que informa o número máximo de condições ativas na regra (*MaxCondAtivas*). Baseando-se nisso, têm-se:

- operador de inserção: é ativado com 2% de probabilidade se o número de condições ativas for menor que *MaxCondAtivas*. Se não houver nenhuma condição ativa, então o operador é executado incondicionalmente (com 100% de probabilidade); e
- operador de remoção: é ativado com 1% de probabilidade caso haja mais de uma condição ativa. Mas, se o número de genes ativos for maior que *MaxCondAtivas*, o operador é executado automaticamente (100% de probabilidade), fazendo a retirada aleatória de condições até que o valor de *MaxCondAtivas* seja atingido.

Existe ainda uma outra situação em que esses operadores podem entrar em ação. Após o *crossover*, se surgirem dois filhos iguais e o número de condições ativas neles for menor que *MaxCondAtivas*/2, executa-se incondicionalmente o operador de inserção para um deles. Se for menor que *MaxCondAtivas*/2, então o operador de remoção é ativado obrigatoriamente em um dos filhos gerados. Esse procedimento procura evitar indivíduos iguais, beneficiando a variedade da população.

6.5 AVALIAÇÃO DAS REGRAS

A avaliação das regras é feita a cada geração de acordo com dois critérios: 1) o cálculo sobre a qualidade da regra; e 2) o cálculo sobre o grau de interesse da regra. Ou seja, para entender a função de *fitness* do algoritmo no AGD é necessário obter o resultado da qualidade e do interesse da regra.

6.5.1 Qualidade da regra

Para poder calcular a taxa de cobertura do atributo difuso, é construída uma matriz, chamada de matriz de confusão (Romão, 2002), conforme demonstra a Tabela 6.4.

Exemplo coberto pelo Antec	Meta Igual	
	SIM	NÃO
Sim	Soma grau de correto (SC) de coberturas "sim"	Soma grau de errado (SE) de coberturas "sim"
Não	Soma grau de errado (NE) de coberturas "não"	Soma grau de correto (NC) de coberturas"não"

Tabela 6.4 - Matriz de confusão difusa

FONTE: ROMÃO, 2002.

Em relação à matriz de confusão, têm-se:

SC (Sim Correto) = antecedente cobre o exemplo, meta igual;

SE (Sim Errado) = antecedente cobre o exemplo, meta diferente;
NE (Não Errado) = antecedente não cobre o exemplo, meta igual;
NC (Não Correto) = antecedente não cobre o exemplo, meta diferente.

Basicamente a matriz apresenta quantitativamente as quatro possíveis alternativas de cobertura, somando os registros que são e os que não são englobados pelo antecedente da regra e cruzando com a informação de se ter previsto corretamente ou não em relação à meta desejada. Através dos resultados obtidos por Romão (2002) no uso de funções que utilizavam as categorias apresentadas na matriz de confusão difusa, foi obtida a seguinte equação, que leva a uma maior taxa de acerto e de cobertura:

$$QUALIDADE = (SC - 1/2) / (SC + SE)$$

O objetivo dessa equação é penalizar regras com baixa cobertura, visto que, com altas taxas de acerto, a subtração de 1/2 não fará diferença relevante no cálculo. É dentro do cálculo da qualidade da regra que o grau de pertinência do antecedente é medido.

6.5.2 Grau de interesse da regra

Enquanto a qualidade da regra pode ser medida apenas através de avaliações objetivas (como cobertura, confiança, simplicidade) (ROMÃO, 2002), medir o quanto uma regra é surpreendente requer também avaliações subjetivas (PIATESTSKY-SHAPIRO, MATHEUS, 1994). Assim, visando aumentar a relevância das regras obtidas e diminuir o número de regras não desejadas pelo usuário, Romão (2002) estudou uma técnica subjetiva de avaliar o interesse existente pelas regras geradas a partir do AGD. Para que essa abordagem seja viável é necessário que o usuário possua conhecimento sobre o domínio do problema, o que neste estudo é verdadeiro.

O conhecimento prévio do usuário é transformado em um conjunto de impressões gerais denominadas IGs, as quais possuem a mesma estrutura da regra. As IGs representam as hipóteses do usuário, seu conhecimento do ambiente do problema. As regras obtidas são confrontadas com as IGs em relação ao grau de similaridade ou diferença entre ambas. O usuário pode estar interessado em dois tipos de resultado para comparação:

f) confirmação de hipóteses: o usuário deseja confirmar sua impressão; ou

g) contradição de hipóteses: o usuário deseja contradizer a impressão que possui.

O grau de similaridade do antecedente (SA) é medido conforme a equação (ROMÃO, 2002) apresentada a seguir.

$$SA_{(i,j)} = \frac{\left|A_{(i,j)}\right|}{\max\left(\left|R_i\right|, \left|IG_j\right|\right)}$$

Onde:

$|R_i|$ = nº de atributos ativos no antecedente da regra R_i descoberta;

$|IG_j|$ = nº de atributos ativos no antecedente da impressão geral IG_j;

$|A_{(i,j)}|$ = nº de atributos ativos de R_i que são iguais (nome e valor) aos atributos ativos da IG_j.

Utilizando o SA, o grau de interesse é calculado considerando o conseqüente contraditório, isto é, comparando IGs e regras com o mesmo atributo conseqüente, em que o antecedente da regra contém pelo menos uma condição igual (em nome do atributo e valor) ao antecedente da IG, mas que possui valor distinto para o atributo meta. O quanto esse valor é distinto contribui para o resultado do cálculo, conforme a distância entre os intervalos das funções de pertinência definidas para o atributo. Como exemplos:

- valor “baixo” na IG e “alto” na regra = grau de interesse máximo;
- valor “alto” na IG e “baixo” na regra = grau de interesse máximo;
- valor “médio” na IG e “alto” na regra = 50% de interesse;
- valor “baixo” na IG e “médio” na regra = 50% de interesse, e assim por diante.

Quando há mais de uma impressão geral para a mesma regra, então o cálculo de grau de interesse obedece à seguinte equação, onde *n* é o número de IGs definidas pelo usuário:

$$INTERESSE = Max(SA_{(i,1)}, SA_{(i,2)}, \ldots, SA_{(i,n)})$$

6.5.3 Função de fitness

Após diversos experimentos utilizando funções que envolviam a qualidade e o grau de interesse da regra, Romão (2002) definiu a seguinte equação para a função de *fitness*:

SE *Interesse > 0*
ENTÃO *Fitness = Qualidade * Interesse*
SENÃO *Fitness = Qualidade / 20*

Essa fórmula objetiva penalizar a qualidade da regra que não possui interesse do usuário. Esse valor (1/20 = 0,05) não é arbitrário, mas foi escolhido para evitar conflitos entre regras sem interesse e com interesse, visto que é impossível uma regra com o mínimo de interesse alcançar 0,05 após multiplicada por qualquer qualidade. Além disso, quando tanto interesse quanto qualidade são maiores que zero, o grau de interesse funciona como um fator depreciador do valor da qualidade (Romão, 2002).

6.6 PARÂMETROS

Vários parâmetros são dados ao AGD quanto aos operadores genéticos, às características próprias de cada amostra e também ao objetivo da mineração de dados. A parametrização é flexível ao tipo de aplicação do AG (item 4.8.4) e é um dos aspectos diferenciais dos algoritmos genéticos (seção 4.11).

No AGD, os parâmetros deixados diretamente no código como constantes do programa são aqueles cujos valores foram encontrados mediante avaliação empírica e se mostraram mais apropriados ao AG, isto é, os aspectos particulares do algoritmo (tipo de método de seleção, tipo de codificação do cromossomo, fórmula da função de payoff, etc.) trabalham conjuntamente em equilíbrio (seção 4.8). Entre os parâmetros próprios do sistema (definidos como constantes) estão os valores para as probabilidades de:

- *crossover* geral;
- *crossover* interno (item 6.4.1);
- mutação geral;

- mutação interna (item 6.4.2);
- inserção e remoção de condições (item 6.4.3);
- ativar um gene categórico;
- ativar um gene difuso.

Outros parâmetros podem ser deixados ao usuário de acordo com o tipo de regra de classificação desejada; portanto, embora eles estivessem declarados no *código fonte*, eles foram modificados aqui para se tornarem variáveis configuráveis pelo usuário. Quanto a esses parâmetros, eles referem-se aos limites máximos de:

- condições ativas na regra;
- gerações;
- tamanho da população;
- tamanho do conjunto de treinamento;
- tamanho do conjunto de testes;
- amostra de dados, a base suporta amostras do mesmo assunto mas de diferentes usuários ou diferentes validades (item 3.3.7).

6.7 SELEÇÃO DA MELHOR REGRA

Quanto à seleção da melhor regra a ser apresentada ao usuário, o algoritmo exige que duas condições sejam preenchidas:

1) Interesse > 0;
2) *Acerto Treinamento* > Max(0,5, *Freqüência Relativa*).

A primeira condição garante que o usuário veja apenas regras para as quais informou algum interesse através das IGs. A segunda condição traduz-se pelo acerto daquela regra durante o treinamento, sendo maior que o máximo entre 0,5 e a freqüência relativa do conseqüente no conjunto de dados. Essa exigência funciona como compensação à facilidade de encontrar tal indivíduo na população, pois, quanto maior a quantidade de registros de uma

determinada classe, mais fácil é prever tal classe. O acerto de treinamento e a freqüência relativa são calculados de acordo com (ROMÃO, 2002):

$$\textit{Acerto Treinamento} = \frac{\text{nº de registros com essa meta/valor classificados corretamente}}{\text{nº de registros com essa meta/valor}}$$

$$\textit{Freqüência Relativa} = \frac{\text{nº de registros com essa meta/valor}}{\text{nº total de registros}}$$

Visto que todos os valores são expressos em porcentagem, 0,5 significará 50%. Desse modo, a condição garante através da função “Max” que as regras apresentadas no final tenham cobertura maior que 50%, mesmo em casos no qual a freqüência relativa da classe seja inferior à metade do conjunto.

6.8 FUNCIONAMENTO DO ALGORITMO

A partir do que foi visto até agora a respeito do AGD, já se pode compreender o funcionamento desse sistema híbrido-difuso. Alguns dos passos citados correspondem às atividades que precisam ser desenvolvidas antes que o programa seja executado, como é o caso, por exemplo, das impressões gerais do usuário. As expressões estão em pseudocódigo, mas o detalhamento lógico das rotinas internas já foi visto nas definições quanto aos operadores genéticos, cálculos da qualidade e do interesse, à função de *payoff* e seleção da melhor regra para apresentação ao usuário. O resumo da arquitetura do algoritmo é apresentado no Quadro 1.

- Obter as IGs do usuário;
- Obter os significados semânticos dos atributos difusos do usuário;
- Repetir para cada par <atributo meta, valor>:
  - Calcular a freqüência relativa do par meta/valor;
  - Formar a população inicial;
  - Computar a qualidade das regras da população inicial;
  - Calcular o grau de interesse das regras da população inicial;
  - Repetir para cada geração:
    - Seleção;
    - Cruzamento;
    - Mutação;
    - Inserção e/ou Remoção de condições nas regras;
    - Computar a qualidade das regras;
    - Calcular o grau de interesse nas regras;
  - Ordenar a população final em ordem descendente de Fitness;
  - Selecionar a melhor regra (maior Fitness) que tenha (Interesse>0) e Acerto de Treinamento > Max (0,5, FreqRelativa);
  - Mostrar a melhor regra da população final para cada meta/valor;
- Fim do algoritmo.

Quadro 1 - Resumo do algoritmo AGD

FONTE: ROMÃO, 2002.

A seguir são apresentadas as razões pela qual o AGD foi sido selecionado como meio para a extração de regras de classificação sobre o específico ambiente de problema deste estudo – as redes de baixa tensão.

6.9 JUSTIFICATIVA

O AGD foi escolhido por possuir aspectos considerados pertinentes ao problema aqui estudado, além de conter características essenciais no que concerne à flexibilidade requerida para bases de dados em constante evolução e compreendendo vários assuntos interessantes ao uso de mineração. No geral, citam-se como fatores decisivos para a seleção desse algoritmo em específico:

– a exploração do conhecimento prévio do usuário através de uma técnica subjetiva, produzindo resultados mais significativos e focalizados no escopo do problema;

- o uso de lógica difusa, melhorando a legibilidade e o entendimento dos resultados sob o ponto de vista dos usuários;
- a possibilidade de troca de condições ativas dentro das regras, permitindo um elevado nível de inovação do genótipo;
- o uso de elitismo, garantindo alto *fitness* sem perda da diversidade;
- a aplicação de metodologia de avaliação inteligente das regras baseando-se não apenas na qualidade mas também no interesse alcançado por essa regras junto aos consumidores de informação;
- a adequação para a aplicação sobre bancos de dados de grande porte;
- o uso de termos lingüísticos como uma discretização natural, reduzindo e simplificando o espaço de busca, o que conseqüentemente otimiza o desempenho do AG;
- a possibilidade de realimentação através de parâmetros informados pelo usuário analista, modificados conforme o conhecimento obtido ao longo do tempo em que o AG tem sido executado.

7 MINERAÇÃO DE DADOS EM REDES DE DISTRIBUIÇÃO DE ENERGIA

Neste capítulo apresentam-se os experimentos realizados com o sistema AGD sobre dados de uma rede elétrica de baixa tensão. A tarefa de extração das regras de classificação envolveu desde a obtenção das amostras de dados e sua preparação até a execução de testes para melhor configurar o AG e seus parâmetros. Além disso, adaptações no algoritmo que se mostraram vantajosas para o seu desempenho e robustez também foram estudadas.

Uma das características propostas por este estudo é que a extração de regras possua relativa autonomia em seu processo para que o algoritmo empregado não requeira um tratamento complexo da amostra de dados, nem tampouco exija que o usuário preencha numerosos metadados ou tenha de estudar profundamente seu funcionamento computacional.

Essa autonomia permitiria o uso direto do AG por engenheiros especialistas, responsáveis por processos de manutenção e planejamento da rede de distribuição. Desse modo, o verdadeiro conhecimento desses especialistas estará direcionado para as características do problema e suas possíveis soluções e não para os aspectos inerentes e específicos da computação evolucionária. Em resumo, o nível de autonomia da solução proposta se refletir-se-á diretamente na aceitação da solução pelo usuário e no direcionamento dos esforços para a qualidade dos resultados gerados do que para o apropriado funcionamento do algoritmo.

Tendo esse objetivo como umas das prioridades, entende-se que uma das principais dificuldades do algoritmo será tratar os dados para seu processamento. No teste simples executado usando um AG sobre redes de baixa tensão (seção 5.2), foram necessárias diversas tarefas analíticas antes que os dados estivessem preparados para a aplicação do algoritmo. Porém, não se pode exigir que o usuário comum, antes de aplicar o AG, primeiramente encontre as correlações entre os atributos de dados disponíveis; também não seria viável exigir que ele divida cada um dos campos numéricos contínuos do banco de dados pela sua distribuição de freqüência e substitua os relativos valores pelas classes de discretização encontradas.

O que se espera do usuário do AG é que ele conheça profundamente o ambiente do problema, o qual se encontra armazenado no *data warehouse*. Nesse caso, para identificar os atributos que lhe interessam e seus relacionamentos, ele conta com a documentação de metadados, as ferramentas OLAP do próprio BD e provavelmente uma pessoa encarregada de gerenciar os dados no banco de dados.

O especialista (ou analista de negócio) é capaz de observar os dados e abstrair informações sem a mesma necessidade de análise exploratória de dados, como o analista técnico (item 3.3.3). Isso ocorre porque se supõe que o especialista tem domínio sobre o escopo do problema, permitindo-se, a partir apenas de seu conhecimento e experiência, fazer inferências e levantar hipóteses direcionadas às necessidades da organização. No caso deste estudo, por serem engenheiros, tais especialistas têm a vantagem de, se desejarem, assumir em parte as atividades de um analista técnico, realizando estatísticas e cálculos complexos para dar suporte às teorias que desejam confirmar.

Embora os especialistas já participem ativamente dos diferentes tipos de processos de extração do conhecimento (item 3.3.2), o papel do minerador de dados – o qual distingue-se do explorador (item 3.3.3) – também é, por natureza, dividido com os usuários da solução que foi encontrada, ou seja, o analista responsável pela mineração utiliza técnicas de *data mining* e implementa o método para testar e validar a veracidade e a força das hipóteses levantadas. Porém, na prática, ainda é o especialista que aprova a solução dada para uso na organização.

O que se propõe é que a extração de regras de classificação, utilizando o algoritmo genético selecionado, seja um processo realizado com o máximo de independência de um analista de sistemas. Em conseqüência disso, pela ampla liberdade de que dispõe o engenheiro, espera-se que ele atue não apenas aplicando sua especialidade na criação de importantes hipóteses mas que também participe ativamente e com interesse pessoal da tarefa de validação dessas hipóteses.

A seguir descrevem-se as atividades relativas à obtenção e manutenção dos dados para o uso do algoritmo genético AGD.

7.1 PREPARAÇÃO DOS DADOS

Para alcançar a autonomia discutida anteriormente, é preciso assegurar que o algoritmo genético será capaz de tratar – até certo nível – problemas que são comumente encontrados em bases de dados, como ausência de valores em determinados campos, tratamento de atributos conforme o tipo (descritivo ou numérico), montagem automática de amostras de teste e de treinamento, aquisição dos metadados necessários conforme um método aperfeiçoado já predefinido, entre outras coisas. Com relação a tais questões, portanto, exige-se que o AG comporte-se flexivelmente de acordo com umas das seguintes opções:

1) dê suporte automático sem quaisquer entradas do usuário; ou
2) forneça ao usuário através de sua interface a chance de decidir o que fazer e/ou como tratar os aspectos específicos dos dados.

As tarefas descritas a seguir com relação aos dados foram desenvolvidas sempre se tendo em mente a necessidade de automação nessa parte do processo. Desse modo, o que não foi possível programar para ser executado ou que não era reconhecidamente simples também não foi implementado no experimento; o objetivo dessa abordagem é fazer com que os resultados sejam obtidos da mesma maneira que a solução proposta obteria se a abordagem fosse usada por um usuário especialista do problema e não por um analista.

É importante observar que deixar ao AG a realização de absolutamente todo o suporte aos dados seria o mesmo que tentar embutir o processo KDD na atividade de mineração, quando, na verdade, é o processo de *data mining* que é interno ao ciclo KDD (item 3.3.1). No entanto, por estar-se trabalhando com dados provenientes de um *data warehouse*, entende-se implicitamente que os detalhes quanto à efetivação da integração dos dados, sua limpeza quanto à inconsistências, a adequada agregação ou requerida granularidade, entre outras tarefas para tratamento da informação, já foram executadas nos processos de ETL, daí a importância de utilizar-se a CIF como fonte para as atividades de mineração de dados (item 3.2.7).

7.1.1 Seleção

É a partir da tarefa de seleção de dados que se pode começar a pôr em prática a metodologia real através do experimento com o AG, isto é, testar como se fará uso do conhecimento dos especialistas. O método adotado para selecionar os atributos pertinentes à amostra, conforme o contexto das regras de classificação pretendidas, baseou-se no processo descrito pelo desenvolvedor do algoritmo, Wesley Romão (2002, p. 190). Através dos experimentos realizados, o autor chegou a uma metodologia para a busca do interesse do usuário e o ajuste do algoritmo para refletir adequadamente esse interesse, conforme resumido no Quadro 2.

- Escolha de um usuário com conhecimento do domínio da aplicação e interesse na mineração de dados.
- Selecionar os atributos de interesse juntamente com o usuário.
- Abstrair IGs a partir de entrevistas com o usuário – uso de questionário – no formato de regras.
- Extrair regras utilizando o algoritmo direcionado pelas IGs.
- Efetuar novas reuniões para apresentação das regras e avaliação do interesse do usuário nessas regras.
- Comparar o grau de interesse informado pelo usuário com aquele fornecido pelo algoritmo.
- Efetuar ajustes no algoritmo e repetir a metodologia.

Quadro 2 - Metodologia para ajuste do algoritmo com relação ao interesse

Embora o objetivo deste trabalho não seja o desenvolvimento do algoritmo genético, a extração de regras de classificação relevantes está diretamente relacionada à adaptação do AGD ao ambiente do problema. Reuniões foram feitas com três usuários especialistas para realizar o levantamento dos aspectos da baixa tensão pertinentes à violação da continuidade no fornecimento de energia elétrica. Além disso, os especialistas também poderiam informar uma abordagem lógica para efetuar-se a divisão racional das amostras, isto é, uma divisão que fizesse sentido no âmbito prático da distribuição de energia.

Antes que a reunião fosse feita, o modelo de dados do *data warehouse* DW Distribuição (item 0) foi analisado. Era preciso escolher o domínio da mineração para poder apresentar algum material aos especialistas e assim coletar seu interesse junto a esse escopo. O *data mart*

“Operação” foi escolhido por estar diretamente envolvido com as informações sobre índices de continuidade. A tabela de fato desse *data mart*, a DESEMPENHO_ATUACAO_EQP, foi selecionada como principal fonte de dados para a extração da amostra, estabelecendo assim o escopo da mineração sobre informações de interrupções no nível de conjunto e não no nível de consumidor.

O fato DESEMPENHO_ATUACAO_EQP relaciona-se com 15 dimensões (uma delas é de controle interno da carga e não está presente), conforme é visto na figura a seguir. A tabela central é o fato, e seus campos cujos nomes são iniciados em “NR_SEQ“ são as chaves estrangeiras, ou seja, as chaves primárias das dimensões que permitem “cortar” (slice) os dados do fato, formando cubos de informação.

O conteúdo dessa tabela de fato – como o próprio nome diz – descreve o desempenho dos equipamentos que atuaram em ocorrências de descontinuidade no fornecimento de energia. Esse fato traz os aspectos das interrupções agregados pelos índices de continuidade de conjunto calculados. É importante observar que nem todas as ocorrências são contadas como descontinuidade (ver item 2.4.1) e as informações dessa tabela sumarizam apenas as que são. Cada uma das dimensões relacionadas é descrita na Tabela 7.5.

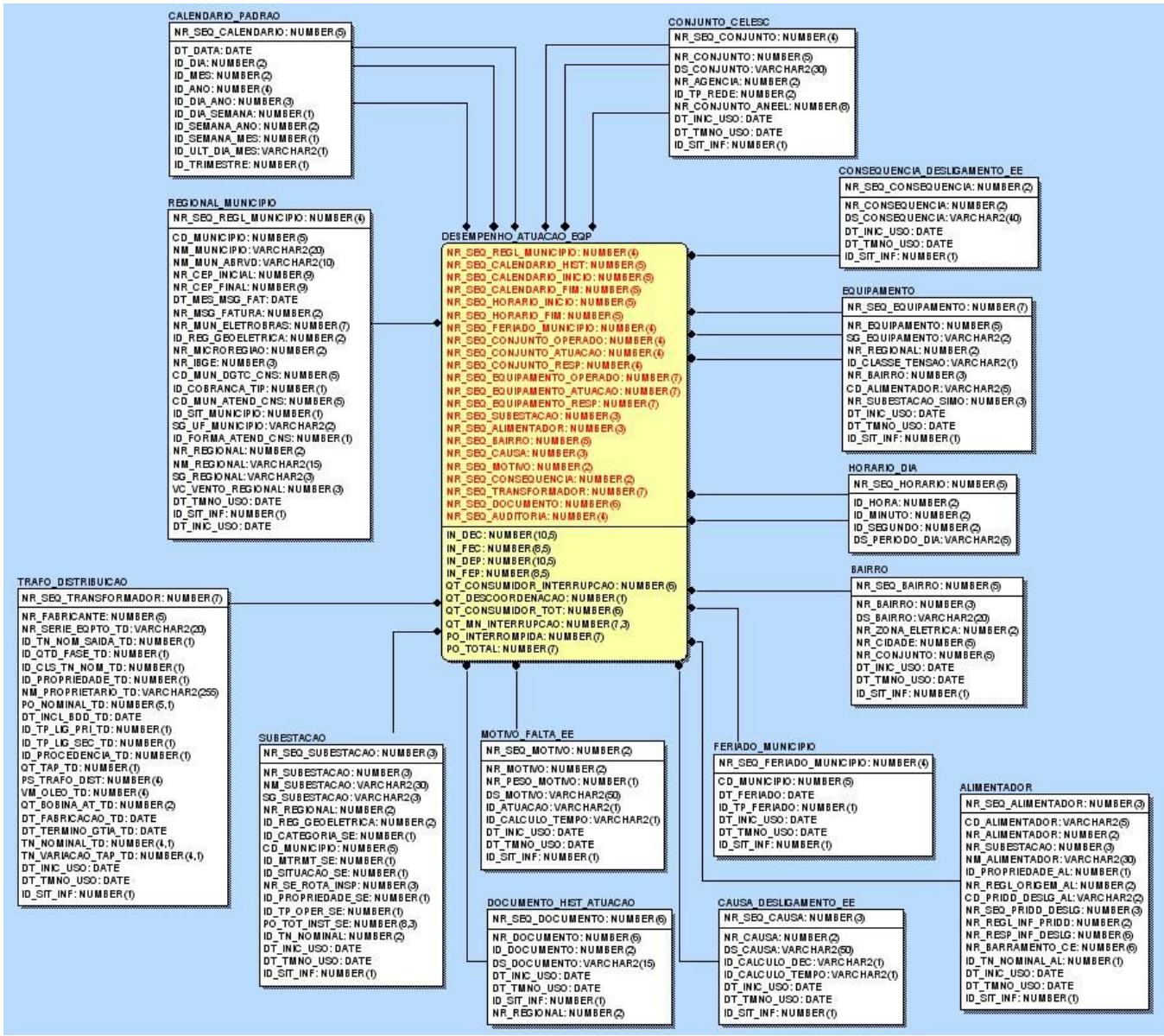

Figura 7.16 - Modelo de dados DW-Distribuição: Fato ATUACAO_EQPTO_REDE_BT

Nome da dimensão	Descrição do conteúdo de dados
ALIMENTADOR	Alimentadores e suas características elétricas e geográficas.
BAIRRO	Bairros separados por zona elétrica, município e conjunto geográfico.
CALENDARIO_PADRAO	Dimensão tempo com grão igual a dia.
CAUSA_DESLIGAMENTO_EE	Causas de interrupção de acordo com avaliação técnica (causas que participam ou não do cálculo de DEC).
CONJUNTO_CELESC	Divisão lógica feita pela CELESC e ANEEL das áreas de distribuição de energia elétrica.
CONSEQUENCIA_DESLIGAMENTO_EE	Conseqüências de interrupção no fornecimento de energia elétrica.
DOCUMENTO_HIST_ATUACAO	Documento de referência à interrupção.
EQUIPAMENTO	Equipamentos e suas informações elétricas e geográficas.
FERIADO_MUNICIPIO	Feriados por município.
HORARIO_DIA	Dimensão tempo com grão igual a segundo.
MOTIVO_FALTA_EE	Motivos de interrupção conforme descritos pelo consumidor que informou a ocorrência.
REGIONAL_MUNICIPIO	Informações geográficas e políticas sobre as agências regionais da CELESC, incluindo seus municípios.
SUBESTACAO	Subestações e suas características elétricas e de operação.
TRAFO_DISTRIBUICAO	Transformadores e seus aspectos físicos, elétricos e de operação.

Tabela 7.5 - Tabelas de dimensão relacionadas ao fato DESEMPENHO_ATUACAO_EQP

Para a primeira reunião foram coletados 57 atributos, dos quais pelo menos um campo de cada tabela foi escolhido. O conjunto de dados foi submetido à análise através da

ferramenta Statistica, observando-se seus indicadores estatísticos básicos (máximo, mínimo, média, mediana e desvio-padrão). Os campos encontrados na base com mais de 60% de valores ausentes ou iguais a zero foram eliminados nessa primeira análise. Em seguida, foi necessário recorrer ao conhecimento dos especialistas para limitar ainda mais o número de atributos relevantes aos índices de continuidade. Durante as entrevistas vários atributos categóricos foram estabelecidos quanto ao seu escopo.

A primeira reunião feita direcionou a lógica da amostragem para refletir a prática em que as redes de distribuição operam no Estado de Santa Catarina. Segundo os especialistas, o comportamento das redes em relação às interrupções no fornecimento está intimamente ligado ao clima e à geografia da região. Assim, concordou-se com a divisão da amostra entre dois grupos de agências regionais: as do litoral e as do interior do Estado. Esses dois grupos geográficos estão descritos na Tabela 7.6.

Regionais do Litoral do Estado	Regionais do Interior do Estado
Criciúma	Blumenau
Florianópolis	Chapecó
Itajaí	Concórdia
Joinville	Jaraguá do Sul
Tubarão	Joaçaba
	Lages
	Mafra
	Rio do Sul
	São Miguel do Oeste
	São Bento do Sul
	Videira

Tabela 7.6 -Agrupamento geográfico para amostragem das agências regionais da CELESC

Já quanto ao clima, os especialistas afirmaram que a adequada sazonalidade deveria ser analisada por mês, visto que mesmo dentro de uma única estação não existe um padrão

conhecido. Com base nisso, em vez de se criarem amostras mensais, o conjunto de dados foi separado por ano e o atributo indicando o mês foi incluído entre os demais, o que permite que cada mês seja analisado dentro do mesmo ano. Desse modo, se houver alguma regra que se caracterize especificamente em um período mensal, ela será encontrada e incluirá o atributo "mês" no seu antecedente.

Após a entrevista, restringiu-se o conjunto de dados a apenas 30 campos interessantes. Com esses atributos, já era possível criar um ambiente para armazenar as informações que seriam utilizadas para a DM: o *exploration warehouse* (ver item 3.2.5). A estrutura do EW para a aplicação do AGD contém basicamente sete tabelas (tabelas comentadas nos próximas seções). Nem todas elas são necessárias, mas todas são úteis. A principal tabela é a de análise, ela é variável quanto a definição de suas colunas porque suas colunas são os atributos da amostra, o que faz com que cada assunto diferente a ser minerado requeira uma outra tabela. No entanto, amostras diferentes sobre o mesmo assunto podem ser armazenadas na mesma estrutura, pois conjuntos diferentes de dados são distinguidos por uma chave estrangeira.

A partir de uma consulta SQL à base, foi criada a tabela de análise contendo esses campos. Por estar inserido no DW Distribuição, o EW provou a eficiência de sua aplicação, facilitando refazer amostras sempre que necessário e também permitindo consultas ao domínio dos atributos de forma automática e integrada durante todo o processo. Por não estar disponível ao usuário, o EW não esteve sujeito à carga de processamento de demais análises, serviu tão somente à mineração, isto é, oferecendo o máximo de desempenho possível.

É importante esclarecer que a inter-relação específica de determinados atributos do conjunto de dados é extremamente indesejada para o uso de uma técnica de previsão, já que pode causar assertivas redundantes. Por exemplo, sabe-se que a quantidade de minutos interrompidos está diretamente ligada ao cálculo de DEC (duração equivalente de interrupção por unidade consumidora), por isso qualquer regra envolvendo ambos os atributos (índice de DEC e quantidade de minutos interrompidos) deveria ser descartada. Apesar desse inconveniente, todos os atributos selecionados foram considerados importantes, e cuidados serão tomados para evitar problemas de redundância nas regras.

Em relação à quantidade de registros, as quatro amostras apresentadas na Tabela 7.7 possuem um número bem distribuído de linhas.

Registros	Litoral	Interior	Total	Litoral %	Interior %
Ano 2004	19.930	52.604	74.136	26,88	73,12
Ano 2005	19.465	49.423	68.888	28,26	71,74

Tabela 7.7 - Número de registros das amostras de dados

7.1.2 Pré-Processamento

Como foi mencionado anteriormente, a devida autonomia do algoritmo genético não pode requerer do usuário qualquer tratamento quanto aos dados. Durante a utilização do AGD por usuários especialistas, espera-se que seu conhecimento sobre as redes de distribuição de energia contribua para que as amostras coletadas tragam apenas atributos relevantes e com domínio restrito a um conteúdo interessante.

Devido à amostragem deste trabalho ter sido realizada pela autora e não por um especialista, foi necessário fazer uso de pelo menos uma ferramenta estatística para melhor entender os dados e seus domínios. Acredita-se que essa prática não será necessária quando o próprio usuário, conhecedor do ambiente de aplicação, estiver montando uma consulta ao banco de dados ou requisitando essa consulta ao responsável técnico pelo BD em questão. Mas, caso um usuário qualquer deseje obter informações específicas da base – como, por exemplo, descobrir se um atributo está sendo preenchido ou não –, uma simples contagem de valores nulos ou distintos no banco de dados irá responder a essas dúvidas.

Durante as entrevistas, buscou-se saber junto aos engenheiros especialistas quais valores – ou intervalo de valores – presentes nos campos coletados do BD deviam ser incluídos no EW. Sendo o objetivo da mineração nesse experimento encontrar regras de classificação que ajudassem a prevenir interrupções no fornecimento de energia assim como auxiliassem na modelagem dos circuitos de baixa tensão, não seria válido estudar a contribuição de ocorrências naturais (aleatórias e fora do controle humano) sobre os índices de continuidade. Por essa razão, o atributo “Causa” foi filtrado para representar apenas fatores passíveis de controle humano. E em seguida o atributo Causa foi confrontado com os índices de DEC e

FEC para avaliarem-se quais as causas que participam mais significativamente do número e da duração das interrupções de energia.

Assim, essa análise de distribuição de freqüência foi feita não apenas para a seleção dos valores mas também para a avaliação da força de influência das causas naturais e das não-naturais sobre a rede. Além disso, pretendia-se investigar o quanto a mineração de dados sobre as causas previsíveis seria relevante considerando o escopo. Os resultados da distribuição de freqüência estão dispostos na Tabela 7.8.

Causas	Naturais	Previsíveis	Total	*Naturais %*	*Previsíveis %*
Quantidade	14	61	75	*18,67*	*81,33*
Número de ocorrências	133.347	117.344	250.691	*53,20*	*46,80*
Valor DEC Acumulado	588.192,95	426.448,88	1.014.591	*57,97*	*42,03*
Valor FEC Acumulado	8.194,59	4.502,97	12.697,55	*64,54*	*35,46*

Tabela 7.8 - Distribuição de freqüência de causas de interrupção nos anos de 2004 e 2005

Através da distribuição de freqüência das causas constatou-se que mais da metade das ocorrências de interrupção era fruto de fatores naturais ou de aspectos impossíveis de serem administrados. Mesmo assim, mediante a representatividade dos valores de DEC e FEC induzidos por causas não-naturais, conclui-se que ainda é altamente significante o número de interrupções por motivos passíveis de previsão. A lista de todas as causas encontra-se no apêndice deste trabalho, na página 159.

Já com o escopo bem definido para esse atributo, efetuou-se uma nova análise de freqüência, desta vez apenas entre as causas previsíveis. A idéia era excluir interrupções que pouco ocorressem na prática, visando deste modo melhorar a eficácia do algoritmo genético ao eliminar-se – ou pelo menos reduzir-se – o genótipo conhecidamente não interessante. É possível observar na Tabela 7.9 o quanto essa análise foi proveitosa, conseguindo excluir

68,85% de causas (42 diferentes causas), as quais contribuíam cumulativamente com apenas 1,52% do DEC e somente 2,04% do FEC gerado pelo total do conjunto de causas previsíveis.

Causas	Existentes	Excluído	Total final	*Excluído %*
Quantidade	61	42	19	*68,85*
Número de ocorrências	117.344	5.926	111.418	*5,05*
Valor DEC Acumulado	426.448,88	6.494,99	419.953,89	*1,52*
Valor FEC Acumulado	4.502,97	92,02	4.410,95	*2,04*

Tabela 7.9 – Distribuição de freqüência das causas previsíveis excluídas da mineração de dados

O conjunto de atributos selecionados ainda passou por mais uma limpeza porque, após feita a divisão de amostras, ele pode assumir valores únicos ou nulos mediante as novas cláusulas, além de tornar evidente campos cujos valores são iguais em todas as tuplas. Por exemplo, um campo que passou a ser preenchido apenas no ano de 2005 possuía somente valores zerados em 2004 e ao dividir-se o conjunto de dados por ano essa discrepância ficou evidente. Um atributo que descreve a classe de tensão no alimentador terá sempre o mesmo valor em relação ao mesmo atributo para o transformador, já que a classe de tensão no final do circuito será necessariamente a mesma da fonte de energia que a alimenta. E ainda um campo que indicava a situação do equipamento, se está operando é igual a 1 e se desativado é igual a 0, assume predominantemente 1 quando a amostra restringe-se apenas a equipamentos que participam de interrupções de energia (equipamentos obviamente em funcionamento).

Essa segunda análise eliminou campos com predominância de valores únicos ou nulos, além de *outliers* para quatro atributos. Como já foi visto neste trabalho (seção 3.3.5), é importante assegurar a não interferência na análise de valores esparsos e acima do limite considerado comum ao escopo do campo (assumem-se valores acima de 4 desvios padrão). Como no experimento pretendido não se buscava encontrar a exceção, mas sim a regra, entendeu-se que os registros retirados da amostra (cerca de 4,5% do total) apenas causariam

problemas à definição dos conjuntos difusos. Assim, a seguir são descritos os atributos e a quantidade respectiva de *outliers* excluídos destes:

- Índice de DEC: 3335.
- Índice de FEC: 127.
- Potência interrompida: 2770.
- Quantidade de minutos interrompidos: 1789.

No final, foram obtidos 13 atributos contínuos e 10 categóricos, num total de 23 atributos participantes da mineração. Suas descrições, tipos e domínio estão na Tabela 7.10.

	Atributos	Tipo	Domínio	Descrição
1	DS_CAUSA	Categórico	-	Descrição da causa de interrupção de energia (avaliação técnica)
2	DS_CONSEQUENCIA	Categórico	-	Descrição da conseqüência de interrupção de energia
3	DS_MOTIVO	Categórico	-	Descrição do motivo de interrupção de energia (consumidor)
4	MTRMT_SE	Categórico	“1” a “3”	Tipo de monitoramento da subestação (“Telecontrolada”, “Telesupervisionada” e “Não Possui”)
5	TP_OPER_SE	Categórico	“1” a “3”	Tipo de observação das operações da subestação (“Assistida”, “Desassistida” e “Parcialmente Assistida”)
6	NM_REGIONAL	Categórico	-	Nome da agência regional em que o alimentador tem origem
7	FASE_TD	Categórico	“1” a “3”	Fases elétricas ligadas ao transformador (“Monofásico”, “Bifásico” e “Trifásico”)
8	PERIODO_INI	Categórico	“1” a “2”	Período do dia em que a interrupção teve início: “Manhã” (0 às 12h) e “Tarde” (12 às 24h)
9	PERIODO_FIM	Categórico	“1” a “2”	Período do dia em que a interrupção terminou: “Manhã” (0 às 12h) e “Tarde” (12 às 24h)
10	ID_TN_NOMINAL_AL	Categórico	“1” a “2”	Tipo da tensão nominal do alimentador (“72,5kV” e “145kV”)
11	ID_MES	Numérico	“1” a “12”	Mês do ano
12	ID_DIA_SEMANA	Numérico	“1” a “7”	Dia da semana
13	IN_DEC	Numérico	“0,0002” a “305,5976”	Índice de DEC
14	IN_FEC	Numérico	“0,00005” a “1,0577”	Índice de FEC
15	QT_CONSUMIDOR_INTERRUPCAO	Numérico	“1” a “4.099”	Quantidade de consumidores interrompidos
16	QT_CONSUMIDOR_TOT	Numérico	“526” a “174.505”	Quantidade de consumidores totais ligados ao transformador
17	QT_MN_INTERRUPCAO	Numérico	“1,667” a “1.789”	Duração em minutos da interrupção
18	PO_INTERROMPIDA	Numérico	“1” a “4.999”	Potência interrompida em Kva
19	PO_TOTAL	Numérico	“1.561” a “720.758”	Potência total no transformador
20	PO_NOMINAL_TD	Numérico	“0” a “500”	Potência nominal do transformador
21	TN_VARIACAO_TAP_TD	Numérico	“0” a “2,4”	Tensão de variação no TAP do transformador
22	HORA_INI	Numérico	“1” a “12”	Hora em que a interrupção teve início
23	HORA_FIM	Numérico	“1” a “12”	Hora em que a interrupção terminou

Tabela 7.10 - Atributos candidatos selecionados

7.1.3 Transformação

A transformação e/ou discretização – atividades comuns para adaptar o conjunto de dados às técnicas que serão aplicadas sobre ele (item 3.3.3) – também não podem ser utilizadas diretamente sobre a amostra coletada devido à complexidade que geraria para usuário.

Assim, para fazer as necessárias modificações nos atributos numéricos, passando-os para os seus respectivos valores categóricos, cada transformação foi efetuada diretamente pela consulta SQL que traz os dados do banco. Para isso, utilizou-se o comando "DECODE" no ORACLE, que permite que valores (inclusive nulos) sejam substituídos por constantes entradas pelo usuário. Basicamente esse comando executa um teste do tipo "SE... ENTÃO... SENÃO...", permitindo também aninhar o próprio comando diversas vezes, produzindo testes como "SE... ENTÃO... SENÃO SE... ENTÃO...", etc. O camndo também pode envolver mais de um atributo. Por exemplo, para que o atributo PERIODO_INI tenha seus valores 0 e 1 transformados em seus verdadeiros significados, o comando inserido na consulta SQL seria:

DECODE (*PERIODO_INI,1, Manhã, Tarde****)***

O comando acima pode ser melhor entendido em pseudo-código como:

SE *PERIODO_INI = 1* ***ENTÃO*** *'Manhã'* ***SENÃO*** *'Tarde'*

As transformações utilizando a própria consulta ao banco de dados tornam o processo de transformação mais simples, inteligível, coeso e reutilizável. Isso ocorre porque ele é feito automaticamente em um único trecho de código, de forma clara e sem correr o risco de não ser executado sobre a amostra, já que está embutido na própria coleta de dados. Os atributos numéricos "categorizados" estão descritos na Tabela 7.11.

Atributo	Descrição	Transformação
MTRMT_SE	Tipo de monitoramento da subestação	1 = Telecontrolada 2 = Telesupervisionada 3 = Sem supervisão
TP_OPER_SE	Tipo de observação das operações da subestação	1 = Assistida 2 = Desassistida 3 = Parcialmente Assistida
FASE_TD	Fases elétricas ligadas ao transformador	1 = Monofásico 2 = Bifásico 3 = Trifásico
PERIODO_INI	Período do dia em que a interrupção teve início	1 = Manhã 2 = Tarde
PERIODO_FIM	Período do dia em que a interrupção terminou	1 = Manhã 2 = Tarde

Tabela 7.11 - Transformações dos atributos numéricos para categóricos

Após as transformações de domínio de informação, ainda restava o adequado tratamento dos dados em relação aos atributos contínuos para uso do AGD. Por incorporar termos linguísticos da linguagem natural e ser capaz de absorver definições concernentes ao ambiente de aplicação (no caso de baixa tensão, por exemplo, nível de tensão “quase crítico”, comprimento do alimentador “longo”, etc.), a lógica difusa oferece naturalmente a discretização e permite flexivelmente o tratamento de incertezas.

O uso de conjuntos difusos é aconselhável em problemas envolvendo dados numéricos em quantidade significativa, como ocorre com a maioria das aplicações de lógica difusa sobre domínios contendo variáveis numéricas contínuas (ROMÃO, 2002). Mas, embora seja uma qualidade significante no escopo dessa aplicação, observa-se, nesse ponto, inerente ao uso da lógica *fuzzy* uma característica complicadora para a solução aqui proposta: a definição dos conjuntos difusos.

Para que o AGD utilize-se de impressões gerais do usuário, ou faça teste da qualidade e do interesse do usuário sobre a regra, além de apresentar as regras obtidas de forma legível, é

necessário fazer uso dos conjuntos difusos, os quais possuem FPs (item 3.3.10) que precisam ser otimizadas.

Romão (2002) argumenta que, para que essa especificação seja a mais próxima possível da realidade, bem como facilmente compreensível pelo usuário, em vez de utilizar-se de um outro algoritmo genético ou de técnicas de busca local programadas, o próprio usuário poderia definir as funções de pertinência. Embora o próprio autor reconheça a perda de generalidade e autonomia nessa abordagem, ele levanta três vantagens para o uso desse procedimento:

1) incorporar conhecimento do usuário sobre o escopo do problema (*background knowledge*);
2) impedir o risco de que o sistema gerasse FPs contra-intuitivas, isto é, domínios que não fizessem sentido, como no exemplo dado por ele (Romão, 2002): Idade até 40 anos considerada “baixa”;
3) reduzir tempo computacional.

Apesar de ser bastante manual prover doze diferentes valores – parâmetros suficientes para definir as três FPs e seus domínios – para cada atributo selecionado para a mineração, isso não representou necessariamente um problema para os seus experimentos relatados (Romão, 2002). No contexto deste estudo, porém, trabalharemos inicialmente (neste experimento) com mais de 15 atributos contínuos, como foi visto no item 7.1.1 desta seção. Se esse experimento for repetido na prática com um especialista comum, seria praticamente inviável exigir que ele entrasse manualmente com 180 valores.

Observando as três vantagens do preenchimento das FPs pelo usuário, chegou-se a uma solução alternativa e flexível. Para preencher esses metadados independentemente do conjunto selecionado para a mineração, foi desenvolvido um procedimento em linguagem PL/SQL. A rotina faz uso das tabelas internas do ORACLE para analisar a quantidade e os tipos de dados das colunas da tabela montada pelo usuário, bem como para calcular valor máximo e mínimo do campo conforme encontrado na amostra e dividir esse intervalo para distribuir os valores entre os conjuntos difusos.

Com essa abordagem, o algoritmo ganha em generalidade, pois aceita praticamente qualquer amostra de dados, também não sofre perda no desempenho computacional, pois esse

procedimento é chamado antes da execução do algoritmo genético. O conhecimento do usuário ainda se faz relevante na medida em que, após o preenchimento automático das funções de pertinência, o especialista pode reduzir o domínio dos conjuntos difusos de acordo com sua experiência no assunto. A solução para a geração de FPs contra-intuitivas também está embutida nessa validação pelo especialista, visto que, se houver disparidade entre os limites calculados (normalmente a disparidade é exceção) e aqueles que corresponderem à realidade, o usuário modificará apenas esses intervalos.

Cabe ressaltar que o sistema AGD já disponibiliza ao usuário uma interface que permite a visualização dos atributos numéricos e seus intervalos, assim como dos atributos categóricos e seus domínios. As FPs otimizadas pelo procedimento PL/SQL também podem ser conferidas no AGD. Desse modo, o usuário tem completo acesso ao conteúdo do banco utilizado pelo algoritmo.

A última transformação necessária sobre o banco de dados antes de algoritmo poder ser testado foi quanto à sua própria estrutura. O programa requer que a tabela de impressões gerais possua os mesmos campos da tabela de análise, porém, todos descritivos – visto que o conteúdo das IGs será categórico ou lingüístico difuso –, além de dois outros campos que armazenam a posição da meta e o seu valor, respectivamente. Dessa maneira, uma mera cópia da definição da tabela de análise para a criação da IG não seria suficiente. Porém, mesmo que o usuário pudesse copiá-la, substituindo os tipos dos campos para textuais, e depois adicionar mais dois campos, o AGD ainda não conseguiria operar corretamente sobre essa tabela, pois o programa exige que os campos descritivos sejam inseridos primeiro e na ordem em que aparecem na tabela de análise, seguidos dos campos numéricos (também em ordem).

Para poupar o usuário desse complicador e não interferir na lógica do algoritmo, foi incorporada no mesmo procedimento PL/SQL uma rotina que lê a tabela de análise, pegando inicialmente os campos descritivos e depois os numéricos, todos ordenados conforme aparecem na tabela de origem e no formato textual. Em seguida, os dois outros campos (próprios da IG) são adicionados, além da chave primária e estrangeira. Por fim, o procedimento certifica-se de que a tabela IG não existe mais no BD – se existir, é excluída – e a recria com a estrutura correta.

O modelo de dados do EW pode ser visto na Figura 7.17. É importante deixar claro que a estrutura das tabelas de análise (DM_REGISTRO_ANALISE) e de impressões gerais (DM_IG) é criada de acordo com o assunto da mineração. Portanto, neste estudo, ambas possuem as colunas selecionadas de acordo com os atributos da amostra, mas vão variar conforme o escopo da mineração de dados.

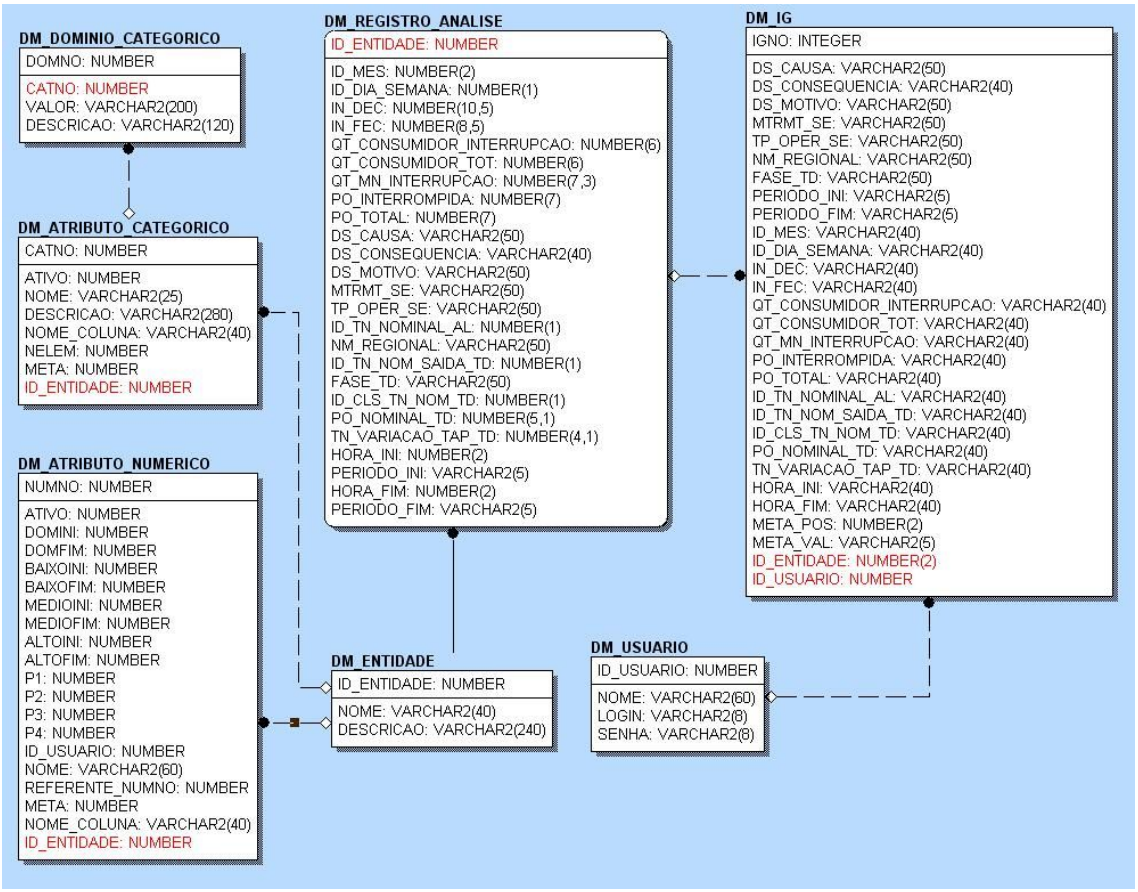

Figura 7.17 - Modelo de dados para o uso do AGD sobre redes de baixa tensão

7.2 MODIFICAÇÕES NO AGD

Para adaptar o sistema AGD às exigências do ambiente de aplicação deste trabalho, diversas mudanças foram feitas no código do programa. Além disso, algumas outras alterações não realmente necessárias também foram implementadas com o objetivo de melhorar os aspectos de desempenho computacional e de interação com o usuário. As modificações são descritas a seguir de acordo com o tipo de alteração.

7.2.1 Interface

Como se pretende que o AGD seja utilizado por usuários para a extração de regras de classificação, foi preciso deixar a interface do sistema o mais simples possível, porém ao mesmo tempo auto-explicativa. Por isso, diversos detalhes foram adicionados ou alterados. As janelas foram padronizadas assim como os componentes de interface utilizados e seus *labels*. Caixas de seleção, *hints* (dicas que aparecem com o passar do mouse sobre o componente), botões, barras de título ativas e atalhos de teclado foram adicionados ou melhor detalhados. O posicionamento dos formulários abertos em relação ao desktop do Windows e o redimensionamento das janelas também foram mudados para não "poluir" a tela, sobrecarregar o usuário com informações ou ainda impedi-lo do uso normal dos demais aplicativos.

Além dessas modificações na estrutura, vários pontos que geravam mensagens de erro ao usuário, ou apresentavam riscos de falha na execução (acesso ao banco de dados, criação de objetos, etc.), foram restritos com cláusulas "try... except" e tratados conforme a necessidade.

7.2.2 Estruturas de dados

Quantas às variáveis, constantes, *arrays*, tipos de dados e demais estruturas de armazenamento de que o programa faz uso, muitos detalhes precisaram ser adaptados ao escopo de dados utilizado por esse experimento, porque, logo que o programa é inicializado, as tabelas do BD são reproduzidas na memória em forma de estruturas de dados. Se estas não forem compatíveis com os domínios de cada atributo, a correta operação do algoritmo é comprometida.

Na análise inicial do código percebeu-se que o algoritmo só dava suporte a valores inteiros no intervalo de 0 a 255. Todas as estruturas de dados e rotinas envolvidas foram alteradas para receber valores contínuos. Pelos testes realizados, o uso de números reais não trouxe problemas para a performance do algoritmo.

A estrutura de dados que mantinha os atributos categóricos armazenava também os diferentes domínios para cada atributo, porém apenas uma palavra em cada posição. Assim, o escopo “Jaraguá do Sul” para o atributo “Nome da Regional” seria carregado parcialmente pelo programa apenas como “Jaraguá”. Esse detalhe também foi alterado, pois havia interesse neste estudo de que os domínios contivessem até mesmo frases, como é o caso das causas e dos motivos de falhas na rede.

Alguns parâmetros fixos no código eram utilizados na declaração de dimensões das matrizes ou dos vetores para definir seus limites máximos. Por exemplo, o intervalo de posições do *array* de atributos numéricos era declarado como sendo de 0 a um valor constante no programa. Se o número de atributos numéricos mudasse no conjunto de análise, o programa lançava exceções de execução ao usuário. Para resolver a inflexibilidade dessas estruturas, todos os *arrays* foram declarados como “abertos”, isto é, sem limite de posições. Isso não é problema para o desempenho do software porque nesse tipo de declaração nenhuma posição de memória é instanciada até que seja necessário, o que ocorre quando a consulta ao banco de dados é feita e o número correto de posições passa a ser conhecido.

Para melhorar a coesão e o encapsulamento de métodos das classes, além da legibilidade do código, a maior parte das variáveis globais passou a ser interna às rotinas que utiliza. Quando isso não foi possível, foram criadas propriedades para as classes, com métodos específicos para atribuição e recebimento dos valores neles contidos.

7.2.3 Entrada e saída de dados

Para carregar os dados em memória, o sistema utilizava determinados componentes que requerem a declaração dos campos em tempo de projeto, precisando armazenar as definições das tabelas que acessavam assim como gerar arquivos em disco cada vez que eram ativados. Visto que o AGD requer que a estrutura das tabelas varie conforme a análise pretendida (item 7.1.1), a definição de campos do BD dentro do programa era bastante contrário à flexibilidade desejada do sistema. Por essa razão, todos esses componentes foram substituídos por outros capazes de aceitar quaisquer atributos existentes na tabela que ele acessa.

7.2.4 Funcionalidade

Uma das modificações mais importantes no AGD foi quanto à habilitação do algoritmo para analisar conseqüentes descritivos. Anteriormente o programa só aceitava atributos contínuos para a meta de uma regra. Isso impedia que quase metade dos atributos da amostra (considerando este estudo) fosse analisada como conseqüentes. Por esse motivo, as rotinas envolvidas nesse aspecto foram alteradas sem que a lógica do sistema mudasse. Para este trabalho, a relevância da nova configuração é refletida no fato de que ela tornou possível analisar causas e conseqüências de falhas elétricas.

Por ser um protótipo desenvolvido especificamente para um determinado caso de uso, o AGD possuía em código as possíveis metas e seus valores correspondentes. Essa característica foi eliminada para dar lugar a uma rotina automática que, a partir das IGs, absorve os diferentes conseqüentes, tanto quanto ao atributo quanto ao valor. Ao ler as IGs da base de dados, cada vez que o programa encontra um novo atributo indicado como meta ou o mesmo porém com valor distinto, o programa considera que o usuário está interessado em uma regra cujo conseqüente contenha esse grupo atributo/valor.

7.2.5 Parametrização

A parametrização do algoritmo era feita completamente através do códigofonte e de alterações diretas nos registros do banco de dados. Por ser um protótipo manipulado apenas por Romão (2002) e somente com o propósito de mineração (somente 1 assunto do qual extrair regras), não era necessário deixar o programa flexível às mudanças de interesse do usuário.

Porém, neste estudo decidiu-se disponibilizar o AGD ao especialista em redes de baixa tensão que desejasse buscar regras de classificação, mesmo que esse usuário não tivesse domínio da ferramenta de programação para efetuar mudanças no código-fonte. Para tanto, alguns poucos parâmetros precisaram ser disponibilizados para modificação através da interface do programa, permitindo ao usuário configurar o algoritmo de acordo com sua amostra de dados e as características do problema. Os seguintes aspectos foram disponibilizados para alteração externa:

1) número de condições ativas na regra;
2) número de gerações;
3) tamanho da população.
4) tamanho do conjunto de treinamento;
5) tamanho do conjunto de testes;
6) probabilidade de ativar um gene categórico;
7) probabilidade de ativar um gene difuso;
8) usuário, senha e *string* de conexão com o banco de dados;
9) amostra de dados do usuário;
10) confirmar ou contradizer IGs.

Em relação ao item 10 é importante acrescentar que o experimento apresentado por Romão (2002) utilizou-se do conseqüente inesperado (seção 3.3.10) visando encontrar maior relevância no conhecimento obtido através do AG. Porém, no experimento deste trabalho em particular, busca-se encontrar regras que orientem a manutenção e o planejamento das redes elétricas de distribuição. Para tal finalidade, é mais simples que as hipóteses levantadas pelos especialistas sejam comparadas em termos de similaridade com as regras. Se as hipóteses não forem verdadeiras, elas serão descartadas, sugerindo que falsas assertivas podem estar sendo aplicadas na prática e por isso merecem ser investigadas.

Apesar disso, acredita-se que outros experimentos possam explorar livremente os demais tipos de comparação entre regra e hipótese (confirmá-las ou contradizê-las), e é por isso que esse parâmetro também foi disponibilizado ao usuário. Inicialmente, implementou-se apenas a verificação do valor do conseqüente, permitindo ao usuário informar ao programa se deseja que as IGs inseridas por ele sejam comparadas em igualdade com o conseqüente ou contrariadas por este.

7.3 APLICAÇÃO DO AGD

Após a preparação de amostras de dados interessantes ao usuário e a execução de modificações no algoritmo para atender às exigências do ambiente do problema, iniciaram-se

os testes para a extração de regras. Durante essa fase do processo, foi preciso definir os conjuntos difusos, obter as IGs dos usuários e setar os parâmetros internos e externos do algoritmo. Essas atividades são descritas a seguir.

7.3.1 Definição dos conjuntos difusos

Para definir as funções de pertinência, foi utilizado o procedimento PL/SQL descrito anteriormente no item 7.1.3. No entanto, como foi mencionado, algumas FPs geradas poderiam ser contra-intuitivas (apenas exceções), cabendo ao usuário a modificação dessas FPs para determinar conjuntos difusos adequados ao atributo. Entre os 23 campos da amostra, apenas o índice de DEC e o índice de FEC precisaram ser modificados quanto às suas quatro funções de pertinência. O critério utilizado foi a distribuição dos valores possíveis dentro do domínio, garantindo assim que os conjuntos difusos possuíssem uma quantidade relativamente igualitária de registros.

7.3.2 Obtenção das impressões gerais

As IGs foram definidas em reuniões com os engenheiros baseando-se em informações fornecidas por eles quanto às intuições que eles têm a respeito dos aspectos relacionados às interrupções na rede de baixa tensão. Neste estudo não foram utilizados formulários como no experimento original (ROMÃO, 2002), mas as hipóteses sobre as possíveis relações entre as características dos circuitos elétricos – no que tange à descontinuidade da distribuição de energia – foram anotadas e inseridas na base de testes.

A importância das impressões gerais do usuário é traduzida pelo papel que desempenha no algoritmo genético. Além das IGs estarem diretamente conectadas à qualidade da regra, é através das impressões gerais que o AG direciona a evolução da população quanto ao número de genes ativos. Também, se o usuário possuir impressões gerais sobre o ambiente do problema muito pontuais, um grande escopo do espaço de soluções é ignorado pelo algoritmo. Por todas essas razões, a definição das IGs tem de ser feita de modo a equilibrar todos os aspectos que envolve.

O número de condições na impressão do usuário não tem influência na qualidade das regras encontradas, pois o grau de qualidade é calculado com base somente nos genes que estejam ativos tanto no indivíduo quanto na IG, enquanto os demais atributos são ignorados. Porém, para o cálculo do interesse, o número de condições é bastante relevante, isso porque o valor do interesse é atingido considerando o grau de similaridade do antecendente da regra em relação às IGs, isto é, quanto maior for a diferença entre o tamanho do antecedente encontrado e aquele informado pelo usuário, menor o valor de interesse naquela regra. Como o *fitness* é calculado sobre o interesse e é a função de *payoff* que determina quais indivíduos sobrevivem, procurou-se manter nas impressões gerais um número intermediário de condições em relação ao tamanho da regra pretendida.

Entende-se que no ambiente deste estudo é importante que o antecedente das regras tenha um número considerável de condições, já que apenas um ou dois atributos dificilmente poderiam descrever a complexidade do comportamento dos circuitos de baixa tensão. A existência de um ou dois aspectos característicos de certa classe de problema (queda de tensão, sobrecarga, etc.) não é surpreendente ao especialista, que, muitas vezes por simples experiência na área, já possui esse conhecimento. Justamente se baseando nesse domínio prévio do problema é que foi escolhida uma abordagem que permitisse aproveitá-lo para direcionar a geração de regras de previsão.

Cada IG foi estipulada relacionando-se com dois ou mais atributos, cujos valores obedeciam ao conhecimento dos especialistas. A determinação desses valores foi relativamente simples, pois a lógica difusa, utilizando-se de variáveis linguísticas, permitiu que as impressões dos usuários pudessem ser traduzidas facilmente.

Cabe ressaltar que foi considerado aqui o problema discutido na página 100 com relação à inter-relação dos atributos no conjunto de dados. Assim as IGs foram restritas quanto ao seu antecedente e conseqüente para minimizar a geração de regras redundantes. Não se pode modificar o algoritmo para tratar desse problema – visto que abrange aspectos específicos de cada análise –, mas neste estudo os campos que possuem correlação direta são conhecidos e qualquer regra envolvendo-os será ignorada.

As IGs levantadas pelos especialistas se referiam aos assuntos a seguir. Independente do assunto, busca-se atingir regras relevantes para os índices de interrupção (DEC e FEC).

Portanto, de algum modo esses índices ou atributos pertencentes ao cálculo dos índices de confiabilidade foram inseridos nas IGs com valores que justificariam interesse, são eles:

a) quantidade de minutos interrompidos = “alto”;
b) potência interrompida = “alto”;
c) quantidade de consumidores interrompidos = “alto”.

1) Interrupções com relação ao período do dia

Acredita-se que a freqüência com que as interrupções ocorrem e a duração dessas interrupções estão relacionadas ao período do dia, visto que o comportamento do tipo de consumidor (seção 2.3) muda conforme o horário. A hora de início ou final não se distinguia nesse contexto bem como o período inicial ou final, por isso, optou-se nesse caso por qualquer um deles.

IG[1]: if PERIODO_INI = Manhã, HORA_INI = baixo, --> Then **IN_DEC = alto**
IG[2]: if PERIODO_INI = Manhã, HORA_INI = medio, --> Then **IN_DEC = alto**
IG[3]: if PERIODO_INI = Manhã, HORA_INI = alto, --> Then **IN_DEC = alto**
IG[4]: if PERIODO_INI = Tarde, HORA_INI = baixo, --> Then **IN_DEC = alto**
IG[5]: if PERIODO_INI = Tarde, HORA_INI = medio, --> Then **IN_DEC = alto**
IG[6]: if PERIODO_INI = Tarde, HORA_INI = alto, --> Then **IN_DEC = alto**
IG[7]: if PERIODO_INI = Manhã, HORA_INI = baixo, --> Then **IN_DEC = medio**
IG[8]: if PERIODO_INI = Manhã, HORA_INI = medio, --> Then **IN_DEC = medio**
IG[9]: if PERIODO_INI = Manhã, HORA_INI = alto, --> Then **IN_DEC = medio**
IG[10]: if PERIODO_INI = Tarde, HORA_INI = baixo, --> Then **IN_DEC = medio**
IG[11]: if PERIODO_INI = Tarde, HORA_INI = medio, --> Then **IN_DEC = medio**
IG[12]: if PERIODO_INI = Tarde, HORA_INI = alto, --> Then **IN_DEC = medio**
IG[13]: if PERIODO_INI = Manhã, HORA_INI = baixo, --> Then **IN_FEC = alto**
IG[14]: if PERIODO_INI = Manhã, HORA_INI = medio, --> Then **IN_FEC = alto**
IG[15]: if PERIODO_INI = Manhã, HORA_INI = alto, --> Then **IN_FEC = alto**
IG[16]: if PERIODO_INI = Tarde, HORA_INI = baixo, --> Then **IN_FEC = alto**
IG[17]: if PERIODO_INI = Tarde, HORA_INI = medio, --> Then **IN_FEC = alto**
IG[18]: if PERIODO_INI = Tarde, HORA_INI = alto, --> Then **IN_FEC = alto**
IG[19]: if PERIODO_INI = Manhã, HORA_INI = baixo, --> Then **IN_FEC = medio**
IG[20]: if PERIODO_INI = Manhã, HORA_INI = medio, --> Then **IN_FEC = medio**
IG[21]: if PERIODO_INI = Manhã, HORA_INI = alto, --> Then **IN_FEC = medio**
IG[22]: if PERIODO_INI = Tarde, HORA_INI = baixo, --> Then **IN_FEC = medio**
IG[23]: if PERIODO_INI = Tarde, HORA_INI = medio, --> Then **IN_FEC = medio**
IG[24]: if PERIODO_INI = Tarde, HORA_INI = alto, --> Then **IN_FEC = medio**

Quadro 3 - IGs sobre interrupções por período do dia

O objetivo era encontrar alguma relação entre o período do dia e os índices de DEC e FEC médios e altos. Os índices de nível médio foram utilizados justamente para poder eliminar pelo complementar os valores baixos, que não interessavam já que são a maioria e não há a intenção de criar um padrão para evitá-los. Assim, as possíveis combinações foram feitas com relação aos índices e às horas dentro dos dois períodos do dia. As IGs relacionadas para esse assunto estão no Quadro 3.

2) Sazonalidade das causas

O primeiro assunto refere-se à hipótese de que algumas causas de interrupções são mais freqüentes durante certos períodos do ano devido a fatores climáticos e populacionais. Regiões que sofrem específica ação da natureza (maresia, geada, alto nível de umidade) e áreas de alta concentração urbana ou de difícil acesso (vilarejos na serra, comunidades rurais, cidades turísticas etc.) podem estar suscetíveis à interrupções por causas diferentes.

IG[1]: if ID_MES = baixo, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[2]: if ID_MES = baixo, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[3]: if ID_MES = alto, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[4]: if ID_MES = baixo, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[5]: if ID_MES = baixo, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[6]: if ID_MES = baixo, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[7]: if ID_MES = alto, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **POSTE AVARIADO, CAIDO, PODRE, OU FORA DE PRUMO**
IG[8]: if ID_MES = alto, PO_INTERROMPIDA = alto,
--> Then DS_CAUSA = **FALHA EM CHAVE FUSÍVEL (FROUXA, MA CONEXAO, OXID)**
IG[9]: if ID_MES = baixo, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[10]: if ID_MES = baixo, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[11]: if ID_MES = alto, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[12]: if ID_MES = baixo, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[13]: if ID_MES = baixo, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **DEFEITO EM PARA-RAIO**
IG[14]: if ID_MES = baixo, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **VEGETAÇÃO NA REDE - MEIO AMBIENTE**
IG[15]: if ID_MES = alto, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **POSTE AVARIADO, CAIDO, PODRE, OU FORA DE PRUMO**
IG[16]: if ID_MES = alto, QT_MN_INTERRUPCAO = alto,
--> Then DS_CAUSA = **FALHA EM CHAVE FUSÍVEL (FROUXA, MA CONEXAO, OXID)**

Quadro 4 - IGs sobre sazonalidade das causas

Considerando esses aspectos, para o começo e fim de ano (mês = “baixo” ou “alto”, respectivamente) foram associadas causas que pudessem estar relacionadas às chuvas (abundantes até março) e ao aumento de consumidores (turistas nos meses de verão). As IG’s relacionadas para esse assunto estão no Quadro 34.

3) Potência interrompida por manutenções programadas[11]

O terceiro ponto de interesse dos usuários era sobre o relacionamento entre a potência dissipada por interrupções programadas. O objetivo era minimizar a potência interrompida pela descontinuidade no fornecimento, causando assim menos prejuízos aos consumidores que possuem alta demanda de eletricidade. Além disso, visto que esse tipo de interrupção contribui largamente para a violação dos índices de continuidade (mais de 34% do DEC total[12] e mais de 29% do FEC total), encontrar um padrão menos custoso com relação à duração da interrupção melhorará os procedimentos para esse tipo de operação sobre a rede, trazendo benefícios significativos para a empresa.

Os quatro tipos de interrupção programada foram combinados com os extremos indesejáveis e desejáveis, em que a quantidade de minutos interrompidos e de potência interrompida é alta no primeiro caso e baixa no segundo. A descrição das impressões gerais para esse assunto é apresentada no Quadro 5.

IG[1]: if DS_CAUSA = PROG. - ALTERAÇÃO PARA AMPLIAÇÃO, QT_MN_INTERRUPCAO = alto, --> Then **PO_INTERROMPIDA = alto**

IG[2]: if DS_CAUSA = PROG. - ALTERAÇÃO PARA MELHORIA, QT_MN_INTERRUPCAO = alto, --> Then **PO_INTERROMPIDA = alto**

IG[3]: if DS_CAUSA = PROG. - MANUTENÇÃO CORRETIVA - EMERGÊNCIA, QT_MN_INTERRUPCAO = alto, --> Then **PO_INTERROMPIDA = alto**

IG[4]: if DS_CAUSA = PROG. - MANUTENÇÃO PREVENTIVA, QT_MN_INTERRUPCAO = alto, --> Then **PO_INTERROMPIDA = alto**

IG[5]: if DS_CAUSA = PROG. - ALTERAÇÃO PARA AMPLIAÇÃO, QT_MN_INTERRUPCAO = baixo, --> Then **PO_INTERROMPIDA = baixo**

IG[6]: if DS_CAUSA = PROG. - ALTERAÇÃO PARA MELHORIA, QT_MN_INTERRUPCAO = baixo, --> Then **PO_INTERROMPIDA = baixo**

IG[7]: if DS_CAUSA = PROG. - MANUTENÇÃO CORRETIVA - EMERGÊNCIA, QT_MN_INTERRUPCAO = baixo, --> Then **PO_INTERROMPIDA = baixo**

IG[8]: if DS_CAUSA = PROG. - MANUTENÇÃO PREVENTIVA, QT_MN_INTERRUPCAO = baixo, --> Then **PO_INTERROMPIDA = baixo**

Quadro 5 - IGs sobre potência interrompida por manutenções programadas

[11] Manutenções previamente determinadas à rede elétrica que causam interrupção no fornecimento de energia aos consumidores.

[12] Valores totais em relação às amostras selecionadas para este estudo.

7.3.3 Parâmetros configurados

Quanto aos parâmetros disponibilizados ao usuário, alguns deles foram mantidos conforme Romão (2002) os havia definido em seu trabalho e os demais foram determinados empiricamente. Por fim, quanto aos parâmetros internos ao programa (seção 6.6), eles não sofreram alteração, pois se considerou que eles foram determinados de forma conjunta para o melhor desempenho do AG de acordo com as características nele implementadas. A seguir estão os valores parametrizados pelo usuário testador.

1) Número de condições ativas na regra: 6.
2) Número de gerações: 30.
3) Tamanho da população: 100.
4) Tamanho do conjunto de treinamento: 2500.
5) Tamanho do conjunto de testes: 2500.
6) Probabilidade de ativar um gene categórico: 10.
7) Probabilidade de ativar um gene difuso: 20
8) Confirmar ou contradizer IGs: Confirmar.

7.4 CONSIDERAÇÕES FINAIS

Ao configurar os parâmetros, a última tarefa necessária para a realização dos testes com o AGD foi finalizada. Foi então possível dar início ao uso do algoritmo genético sobre os assuntos anteriormente citados. O sistema foi executado em média 5 (cinco) vezes para cada conjunto de IGs, e cada conjunto de impressões gerais foi aplicado separadamente para cada amostra de dados – uma relativa ao litoral e outra ao interior. As descobertas feitas, as limitações encontradas, as vantagens do uso dessa ferramenta no ambiente deste estudo e as regras de classificação obtidas durante a aplicação dos testes são descritas e discutidas no capítulo a seguir.

8 RESULTADOS E DISCUSSÃO

Neste capítulo as regras de classificação encontradas são comparadas com outros métodos comumente utilizados referentes à eficiência, às limitações, à autonomia e à interação com o usuário. E, principalmente, são considerados para efeitos de avaliação a exatidão e o grau de relevância e de compreensão das regras encontradas.

8.1 REGRAS DE CLASSIFICAÇÃO OBTIDAS

Mais de 60 impressões gerais foram passadas ao AG, e o processamento sobre elas gerou mais de 50 diferentes regras, entre as quais foram selecionadas as duas melhores para cada assunto abordado. O critério utilizado para essa seleção se baseou nos valores para a cobertura, freqüência relativa e taxa de acerto da regra. Os dois últimos indicadores já foram descritos na seção 6.7, mas, para facilitar o entendimento do processo efetuado neste capítulo, esses indicadores são apresentados novamente a seguir.

$$\textit{Cobertura} = \frac{\text{nº de registros cobertos pelo antecedente}}{\text{nº total de registros}}$$

$$\textit{Freqüência Relativa} = \frac{\text{nº de registros com essa meta/valor}}{\text{nº total de registros}}$$

$$\textit{Acerto} = \frac{\text{nº de registros com essa meta/valor classificados corretamente}}{\text{nº de registros com essa meta/valor}}$$

As regras selecionadas para discussão são apresentadas a seguir de acordo com o assunto a que se referiam. Os indicadores são divididos por litoral e interior, mas, por possuírem valores relativamente similares, restringiu-se a comentar apenas os indicadores relacionados às regras obtidas para a região litorânea.

8.1.1 Interrupções com relação ao período do dia

Para este ponto de interesse parecia fácil achar um padrão, pois a intuição de que em certos horários do dia o consumo de eletricidade possui picos é um consenso entre os especialistas. Porém, permanece complexo achar um modelo geral, já que o tipo de consumidor influi diretamente sobre o período do dia, e vários aspectos a isso relacionados estão dispersos entre os registros. Por exemplo, sabe-se que consumidores industriais vão exigir alta potência durante o período diurno, muito mais do que as áreas residenciais durante a noite, porém, a demanda comercial e industrial é relativamente bem distribuída e contínua, enquanto a residencial ocorre mais aleatoriamente e pode possuir intensidade acumulada num mesmo intervalo (em um determinado dia de muito calor, uma grande quantidade de pessoas decide chegar em casa – o que ocorre por volta do mesmo horário – e ligar seus aparelhos de ar-condicionado).

A primeira regra encontrada (Tabela 8.12) para esse assunto foi obtida a partir de outras duas regras. Notando que apareciam indivíduos na população com alto índice de DEC trazendo o horário aproximadamente a partir do meio da manhã (quando as atividades do dia se iniciam), essas duas regras foram unidas, criando um complementar: horário de início <> “baixo”, isto é, “médio” e “alto”. Como essa abordagem possui lógica, permitiu-se utilizá-la aqui. O resultado foi uma regra com ampla cobertura e alta taxa de acerto.

Regra 1	IF (ID_TN_NOMINAL_AL = 72,5kV) (PERIODO_INI = Manhã) (HORA_INI <> baixo) → THEN (**IN_DEC = alto**)					
	Em transformadores alimentados por tensão igual a 72,5 kV, no período da manhã, das quatro horas ao meio-dia → **DEC >= 0,8**					
Litoral	**Cobertura:**	56,55%	**F. Relativa:**	17,8%	**Acerto:**	48,11%
Interior	**Cobertura:**	25,38%	**F. Relativa:**	25,93%	**Acerto:**	16,73%

Tabela 8.12 - Interrupções com relação ao período do dia: Regra 1

A regra de classificação é simples e traduz-se na afirmação de que transformadores alimentados por média tensão tendem a sofrer interrupções com até 19,9 para o valor de DEC e causar descontinuidades que podem durar horas. Ao alcançar quase 50% de registros corretamente classificados, acredita-se que essa regra possa ser utilizada em estudos sobre a

correlação da tensão de entrada proveniente da rede de transmissão em relação às interrupções diurnas nas redes de distribuição. Ou ainda, essa regra pode ajudar a definir o balanceamento de carregamento dos circuitos a partir da análise de atividades na área urbana durante o horário comercial.

Para efeito de validação, investigou-se a quantidade de transformadores na amostra de dados que são alimentados por média tensão e os que são abastecidos por alta-tensão. Pretendia-se descobrir se a regra era tendenciosa à medida que uma das classes de tensão era predominante. Porém, a diferença quanto à distribuição de freqüência desses dois atributos não foi achada significativamente grande para causar uma tendência. Cabe observar nesse detalhe a importância do conhecimento do especialista durante a análise, pois um engenheiro trabalhando com redes de distribuição no Estado poderia informar o aspecto acima investigado por simples intimidade em sua área de trabalho.

A partir da descoberta dessa característica, mais aprofundamentos foram feitos nessa direção, desta vez procurando relacionar o tipo de monitoramento na subestação com o índice de DEC ou FEC. Mas antes de selecionar a segunda regra, ela foi comparada com outra que também apresentava as mesmas condições (atributo e valor) e altos indicadores. A diferença estava no conseqüente, que em vez de “alto” era “médio”. A cobertura de ambas era obviamente igual e a taxa de acerto diferia de apenas 42,05% para 46,22%. O fator decisivo foi a freqüência relativa de ambas, 17,8% para a meta igual a “alto” e 36,92% para a meta igual a “médio”, o que significa dizer que o segundo conseqüente ocorre duas vezes mais que o primeiro na população e, portanto, caracteriza mais eficazmente um comportamento.

Regra 2	IF (MTRMT_SE = Não Possui) (PERIODO_INI = Manhã) (HORA_INI <> baixo) → THEN (**IN_DEC = medio**)					
	Em transformadores que não possuem monitoramento na subestação, no período da manhã, das quatro horas ao meio-dia → **0,79 >= DEC >= 0,1**					
Litoral	**Cobertura:**	49,16%	**F. Relativa:**	36,92%	**Acerto:**	46,22%
Interior	**Cobertura:**	44,10%	**F. Relativa:**	37,58%	**Acerto:**	41,87%

Tabela 8.13 - Interrupções com relação ao período do dia: Regra 2

A regra selecionada confirma a suposição de que subestações não monitoradas possuem circuitos ligados a ela mais problemáticos, como pode ser visto na Tabela 8.13. A distribuição de freqüência dos tipos de monitoramento apresentava-se bastante balanceada (cerca de 60% das subestações no litoral e 40% no interior não possuem qualquer tipo de monitoramento).

No entanto, o tipo de monitoramento em andamento na subestação não é um aspecto cuja modificação requeira investimentos tão dispendiosos (senão até mesmo fora de cogitação) quanto mudar o tipo de alimentação dos circuitos de baixa tensão (Regra 1 deste assunto). É justamente visando encontrar justificativas – em termos de valores – para mudanças passíveis de serem implementadas que este estudo envolve aspectos como os procedimentos em execução para o controle da distribuição de energia (tipo da operação e tipo de monitoramento na subestação). Nenhuma regra foi encontrada relacionando o período do dia e a alta freqüência de interrupções. Confirmar isso não era exatamente esperado pelo usuário, mas se acreditou válido investigar.

8.1.2 Sazonalidade das causas

Em relação a esse assunto não foram encontradas regras com taxa de acerto relevante. Quanto ao significado da “Potência total” no contexto da interrupção, este campo se refere à potência em kVA fornecida à área urbana que sofreu interrupção; portanto, ele pode referenciar-se tanto à classe de consumidor de uma região quanto ao número de consumidores atendidos por esse equipamento. Por exemplo, uma área pequena onde estão instaladas indústrias ou empresas comerciais terá alta demanda de potência no transformador, mas isso não pode ser afirmado, já que grandes áreas residenciais com alto índice demográfico também podem possuir a mesma característica.

A primeira regra (Tabela 8.14) pode ser mais ainda abstraída para traduzir-se na afirmação de que no fim do ano, em áreas rurais ou pequenas áreas residenciais, podem ocorrer interrupções de energia por causa de vegetação da rede elétrica durante o fim de semana. A relação dos dias da semana com o período anual em que a causa ocorre não tem sentido lógico, o que também pode ser dito em relação a essa causa específica – pode ser mera coincidência que a vegetação interfira na rede elétrica apenas em certos dias da semana. Por

outro lado, é coerente a idéia de que áreas residenciais ou rurais sofram mais com o crescimento da vegetação durante certos meses do que centros industriais e comerciais.

Regra 1	SE (ID_MES = baixo) (ID_DIA_SEMANA = alto) (PO_TOTAL = baixo) → ENTÃO (**DS_CAUSA = VEGETAÇÃO NA REDE - MEIO AMBIENTE**)					
	Meses entre outubro e dezembro, nos últimos dias da semana (de quinta a sábado), com potência total no transformador <= 245.650kVA → **VEGETAÇÃO NA REDE - MEIO AMBIENTE**					
Litoral	**Cobertura:**	13,60%	**F. Relativa:**	9,25%	**Acerto:**	13,68%
Interior	**Cobertura:**	16,32%	**F. Relativa:**	15,43%	**Acerto:**	17,41%

Tabela 8.14 - Sazonalidade das causas: Regra 1

Assim, embora essa regra tenha alcançado um nível de interesse relevante para o usuário, pois relacionou atributos mencionados nas suas impressões gerais, sua taxa de acerto deixa dúvidas quanto à sua utilidade e caberia mais como conhecimento complementar numa análise sobre esse assunto do que realmente como um padrão relacionando essa causa à época do ano.

Entre as duas melhores regras, a segunda (Tabela 8.15) alcança maior índice de acerto, além de ser mais simples – possui apenas duas condições. Devido à mutação ocorrida durante a evolução da população, essa regra faz referência a meses do meio do ano (não constava nas IGs fornecidas) e pode ser interpretada intuitivamente como: “durante o inverno, o consumo de energia é elevadíssimo durante os últimos dias da semana chegando a causar sobrecarga no transformador”.

Regra 2	SE (ID_MES = medio) (ID_DIA_SEMANA = alto) --> ENTÃO (**DS_CAUSA = SOBRECARGA NO TRANSFORMADOR**)					
	Meses entre maio e julho, nos últimos dias das semana (de quinta a sábado) → **SOBRECARGA NO TRANSFORMADOR**					
Litoral	**Cobertura:**	10,66%	**F. Relativa:**	3,71%	**Acerto:**	20,88%
Interior	**Cobertura:**	10,78%	**F. Relativa:**	2,46%	**Acerto:**	19,91%

Tabela 8.15 - Sazonalidade das causas: Regra 2

A cobertura da regra indica que esse antecedente abrange mais de 10% dos registros, mas apenas 3,71% de todas as interrupções são geradas por essa causa. Isso quer dizer que caracterizá-la não seria uma tarefa simples, principalmente quando há tantas variáveis envolvidas. Mais uma vez a baixa taxa de acerto não garante um padrão para a utilização, apenas levanta pontos a serem mais bem investigados pelo especialista em novas IGs sobre o assunto.

8.1.3 Potência interrompida por manutenções programadas

Para esta regra houve predominância de duas causas de acordo com o valor da meta. Para pouca interrupção de demanda durante manutenções programadas, o AG encontrou apenas regras com a causa “PROG. - ALTERAÇÃO PARA MELHORIA”. Sobre as IGs cujo conseqüente era baixa interrução de potência a única causa apresentada nas regras foi a “PROG. - MANUTENÇÃO PREVENTIVA”.

Além da baixa taxa de acerto e conter apenas duas condições no antecedente, a freqüência relativa da regra 1 (Tabela 8.16) desmotiva seu uso para descrever qualquer padrão entre os dados. No entanto, entende-se que por serem previstas, tais interrupções já são organizadas na tentativa de diminuir o máximo possível a potência que seja dissipada durante as modificações necessárias executadas na rede.

Regra 1	IF (DS_CAUSA = PROG. - ALTERAÇÃO PARA MELHORIA) (PERIODO_INI = Manhã) → THEN (PO_INTERROMPIDA = alto)					
	Durante o período da meia-noite ao meio-dia, alterações programadas para melhoria do circuito → **PO_INTERROMPIDA > 3333kVA → 0,79 >= DEC >= 0,1**					
Litoral	**Cobertura:**	14,96%	**F. Relativa:**	0,8%	**Acerto:**	15,09%
Interior	**Cobertura:**	12,18%	**F. Relativa:**	0,94%	**Acerto:**	9,86%

Tabela 8.16 - Potência interrompida por manutenções programadas: Regra 1

A segunda regra encontrada buscava descobrir um modelo ideal para a execução da manutenção, em que o menor valor possível de potência interrompida na manobra fosse

atingido. Mas, além disso, outros fatores importantes estão relacionados e podem ser relativamente controlados, como, por exemplo, o número de consumidores atingidos pela interrupção (influenciando diretamente a violação de índices individuais de continuidade, descritos no item 2.4.1). Assim, a regra que conseguiu reunir as características mais otimizadas foi selecionada e é descrita na Tabela 8.17.

Regra 2	IF (QT_CONSUMIDOR_INTERRUPCAO = baixo) (QT_MN_INTERRUPCAO = baixo) (DS_CAUSA = PROG. - MANUTENÇÃO PREVENTIVA') (PERIODO_INI = Manhã) → THEN (**PO_INTERROMPIDA = baixo**)					
	Durante o período da meia noite ao meio dia, alterações programadas para manutenção preventiva, afetam menos de 1353 consumidores, duram menos de 591,18 minutos → **PO_INTERROMPIDA < 1650kVA**					
Litoral	**Cobertura:**	7,96%	**F. Relativa:**	97,6%	**Acerto:**	8,16%
Interior	**Cobertura:**	5,99%	**F. Relativa:**	96,87%	**Acerto:**	6,12%

Tabela 8.17 - Potência interrompida por manutenções programadas: Regra 2

Mais uma vez a alta freqüência dessa característica para a meta (atributo e valor) dificultou a chance de encontrar um padrão em um meio tão diversificado de valores para os demais atributos da base. Embora esse indivíduo não possa ser utilizado como modelo, ele pode descrever quais as ideais características de uma manutenção programada, mostrando que tal conjunto de aspectos já pôde ser aplicado antes.

8.2 OBSERVAÇÕES GERAIS

Com relação ao litoral ou ao interior do Estado, ao contrário do que se esperava, não foram obtidas regras de classificação distintas, mas apenas as mesmas regras com diferentes valores para seus respectivos indicadores. Pelo que pôde ser observado, a cobertura, a freqüência relativa e a taxa de acerto não se distinguiram significativamente para que o AGD seja aplicado em amostras separadas novamente. De qualquer modo, sem que esse experimento fosse realizado, ainda restavam muitas dúvidas sobre o quanto as características das redes de distribuição divergem ao comparar-se a região do interior e do litoral do Estado.

Muitas regras consideradas pelo algoritmo como sendo interessantes para o usuário não foram apresentadas nesse experimento pela amplitude já determinada do estudo aqui proposto. Além disso, aprofundamentos quanto a pontos de interesse poderiam ser efetuados através da realimentação das IGs com novas hipóteses em vista após os primeiros testes, porém, isso iria requerer muito mais interações com os especialistas para receber o *feedback* deles. Em vez disso, o objetivo dos testes iniciais – além de estudar o potencial da ferramenta – consistiu em estimular os usuários do AGD aplicando conhecimento prévio dos problemas já conhecidos e buscando pontos que mais atendessem às suas necessidades no dia-a-dia.

As regras de classificação obtidas não são de modo algum definitivas quanto ao seu conteúdo, nem garantem resultados já validados. A própria característica evolucionária da abordagem deixa em aberto o limite de possibilidades para investigações mais detalhadas dos assuntos aqui citados, além de muitos outros que tenham relevância no ambiente de baixa tensão. A validade das amostras por si só já reflete a dinâmica a que os dados da rede de baixa tensão estão sujeitos e induz à contínua busca por regras atualizadas bem como por novas descobertas e aplicações.

Finalmente, a utilização do AGD por diferentes especialistas tende a enriquecer o escopo das possíveis análises, à medida que, ao inserir mais experiência quanto aos problemas existentes, amplia o espaço de buscas e abre caminho para a extração de conhecimento novo, útil e aplicável.

8.3 ANÁLISE DOS RESULTADOS

Para fazer a análise dos resultados de modo tangível é necessário quantificar os resultados obtidos de forma que eles possam ser expressos em valores. Nesse caso, há duas possíveis transformações que podem ser derivadas do uso das regras de classificação geradas.

A primeira forma de quantificar os benefícios recebidos é feita calculando-se o quanto seria reduzido dos atuais índices de continuidade em termos percentuais. Os valores obtidos dariam uma idéia da verdadeira viabilidade da aplicação das regras e de quanto poderia ser economizado em relação a multas por violação no fornecimento de energia.

O segundo método para medir o ganho com as regras de classificação geradas baseia-se em um trabalho desenvolvido pelos engenheiros da CELESC (CELESC, dez. 2004) para medir a "energia não-distribuída" (END). O princípio é de que, ao interromper o fornecimento de energia elétrica, o cliente deixa de consumir e por isso também deixa de ser cobrado durante o período da interrupção. Certos tipos de usuário podem ter apenas postergado a atividade que consumiria eletricidade (por exemplo, uma pessoa deixa para tomar banho quando a energia elétrica for restabelecida), mas o tipo de consumidor que não pode adiar as atividades é o que mais contribui para a demanda elétrica (clientes industriais e comerciais, que algumas vezes possuem transformadores dedicados somente ao seu negócio). No estudo realizado (CELESC, dez. 2004), para quantificar a energia não-distribuída devido às interrupções no fornecimento, foi realizado o seguinte cálculo:

$$END = \frac{(\text{Potência Interrompida} * \text{Quantidade de minutos interrompidos})}{60 \text{ minutos}}$$

A unidade da END é o KWh. Para saber o quanto a empresa distribuidora de energia deixaria de arrecadar com a END, chegou-se à fórmula a seguir.

$$\text{Perda de receita} = END * \text{tarifa cobrada por quilowatt/hora}$$

Na fórmula a tarifa citada acima é de R$ 0,40 no Estado. Além da receita, foi computado o custo social envolvido na interrupção, isto é, qual o prejuízo para a sociedade em termos monetários por causa da descontinuidade no fornecimento de energia. Empresas deixam de produzir ou tem sua produção comprometida se durante seus horários de operação elas ficam impossibilitadas de trabalhar (CELESC, dez. de 2004). A pesquisa sobre o custo social associado a interrupções de energia elétrica foi feita por Freire (1999) e chegou-se ao valor de R$ 2,40 centavos por KWh.

Para essa transformação, as regras foram aplicadas em forma de consultas SQL sobre todas as amostras de dados integradas em um único conjunto. O cálculo da perda de receita foi realizado automaticamente pelo banco de dados[13] – sem interferência do analista – para estimar o lucro que seria obtido pela empresa a partir da energia não-distribuída devido a

[13] Os valores para potência e minutos interrompidos foram selecionados como colunas da consulta SQL e multiplicados ou divididos (conforme a necessidade) na própria consulta SQL – isto é, pelo banco de dados – pelas constantes referentes à tarifa, ao custo social, ao número de minutos, etc.

interrupções no fornecimento de eletricidade (ver seção 11.2). Considerou-se a hipótese de que as regras de classificação descritas neste trabalho fossem aplicadas em apenas 10% dos registros que elas classificaram. Desse modo, a Tabela 8.18 descreve a receita que a companhia deixaria de perder e os índices de DEC e FEC que evitaria se houvesse conseguido reduzir em 10% os circuitos problemáticos mapeados pelo conseqüente das regras encontradas pelo AGD.

É importante ressaltar que, embora a amostra contenha dois anos de dados sobre todo o Estado de Santa Catarina, os valores foram divididos pela metade para tornar mais compreensível o lucro estimado, limitando-o por ano. A hipótese de redução de 10% dos circuitos detectados com problema ou passíveis de serem otimizados também é válida para os valores de DEC e FEC. Por exemplo, para a Regra 1 do assunto “interrupções por período do dia”, estima-se que a companhia deixou de receber anualmente em média cerca de R$ 259.274,41 por causa de interrupções, as quais tiveram um custo social de aproximadamente R$ 1.555.646,46, além de somar 280,42 ao seu índice de DEC e 3,43 ao índice de FEC.

Assuntos	Regras	Perda de Receita R$	Custo Social R$	DEC	FEC
1. Interrupções por Período do dia	Regra 1	259.274,41	1.555.646,46	1.470,01	13,70
	Regra 2	123.817,53	742.905,22	346,23	4,83
2. Sazonalidade das Causas	Regra 1	28.442,74	170.656,47	280,42	3,43
	Regra 2	1.097,61	6.585,66	11,27	0,13
3. Potência por Manutenções	Regra 1	30.994,67	185.968,04	30,46	0,95
	Regra 2	50.606,99	303.641,94	426,60	3,13
Total		**494.233,45**	**2.965.404,06**	**2.564,99**	**26,17**

Tabela 8.18 - Perda de receita anual gerada por END e respectivos DEC e FEC causados

Os valores para os índices de continuidade não podem ser diretamente computados em valores monetários porque existem metas a serem cumpridas, isto é, somente quando são violadas as metas estipuladas pela ANEEL naquele mês, trimestre ou ano, então esses valores se tornam multas aplicadas à companhia concessionária de energia.

Observa-se após essa análise que as regras que tiveram pouco suporte (Regra 2 do assunto 2 e Regra 1 do assunto 3) não atingiram valores significantes, induzindo a crer que a cobertura da regra está relacionada em geral à relevância dos resultados transformados dessa classificação.

8.4 CONSIDERAÇÕES FINAIS

A análise dos resultados através da transformação do conseqüente das regras em valores tangíveis mostrou que as regras de classificação encontradas pelo AGD possuem grande potencial para limitar as perdas de receita da companhia distribuidora causadas por interrupções no fornecimento, ao mesmo tempo que reduz os índices de interruções monitorados pela ANEEL. Ao se aplicar o conhecimento encontrado pela mineração de dados em apenas 10% dos circuitos de baixa tensão classificados pelo AG de acordo com o interesse do usuário, constatou-se uma perda de receita relativamente alta.

Cabem aqui algumas considerações a respeito do uso do algoritmo genético durante esse experimento. Através dos resultados obtidos provou-se a capacidade do AGD de gerar regras de classificação de qualidade. Sua utilização tornou evidente muitos benefícios que eram desconhecidos quando esse AG foi selecionado, mas também descobriu-se algumas pequenas desvantagens em sua aplicação no ambiente de problema aqui descrito.

Em relação aos aspectos negativos encontrados, percebeu-se que quando há a predominância de um determinado valor na população, ocorre uma rápida convergência para um mesmo indivíduo, diminuindo a possibilidade de inovação genética. O contrário, isto é, quando os valores buscados são exceção na população, também se mostrou um problema, pois tais indivíduos, embora gerados pelo AG, dificilmente eram selecionados para a apresentação ao usuário.

Entre as vantagens do AGD encontradas a partir dos testes, conta-se a capacidade do algoritmo de atingir regras que contrariem às metas (com relação ao valor) definidas nas IGs. Caso seja gerado um indivíduo com alto *fitness* e cujo valor do conseqüente contrarie o

interesse predefinido do usuário, esse indivíduo não perde a chance de ser selecionado entre as melhores regras, deixando assim ao analista a decisão de considerá-lo relevante.

Outra característica positiva e inerente à ferramenta é a o fator de inovação de conteúdo genético que a mutação acrescenta à técnica. Através do operador de mutação se pôde chegar a valores de atributos que não necessariamente existiam na base, mas que configuravam valores ótimos para o objetivo pretendido.

Mais um ponto que se mostrou vantajoso quanto a esse sistema híbrido-difuso foi a possibilidade de trabalhar com metadados para o processamento do conjunto de registros. As tabelas utilizadas pelo sistema flexibilizam a manipulação de valores semânticos, permitem a desabilitação temporária de atributos e oferecem a chance de se trabalhar com amostras de dados de diferentes usuários e com validades distintas.

Por último e mais importante, a característica do AGD de orientar a análise de acordo com o interesse do usuário tornou o processo de descoberta das regras de classificação bastante eficiente, simples, prático e compreensível ao usuário analista. Enquanto o J4.8, aplicado sobre à classe “Causa” (com 19 possíveis valores para o atributo) usando a ferramenta Weka, gerou 8135 folhas na árvore de decisão construída e um total de 19930 instâncias, o AGD produziu apenas 1 ótima regra para a mesma classe (atributo/valor) em cada execução realizada – em média apenas 5 execuções foram necessárias antes que uma regra considerada como pertinente, válida e significativa fosse selecionada pelo usuário.

Além de não sobrecarregar o analista com regras que não lhe interessam, a ferramenta também elimina naturalmente da população os atributos que não estão envolvidos com as impressões gerais.

Não cabe no escopo deste trabalho analisar a viabilidade e pertinência das possíveis modificações a serem tomadas para corrigir falhas na rede de distribuição e assim solucionar os pontos evidenciados pelas regras obtidas. Porém, espera-se que os testes realizados neste estudo possam demonstrar com eficácia a capacidade da abordagem evolucionária em conjunto com técnicas de mineração de dados para extrair conhecimento válido, surpreendente, compreensível, interessante e útil ao usuário analista.

9 CONCLUSÕES

Este trabalho descreveu uma aplicação de mineração de dados em redes de baixa tensão utilizando algoritmos genéticos. Durante este estudo, foram levantadas as características gerais dos sistemas elétricos, os conceitos de *data warehouse*, fábrica de informações, *data mining* e algoritmos genéticos, além de terem sido focalizadas as possíveis técnicas de mineração e a adequação da aplicação de AGs no escopo de redes de distribuição elétrica.

A pesquisa se desenvolveu no ambiente de uma empresa concessionária de energia elétrica, a CELESC. Inicialmente, procurou-se identificar suas possíveis demandas por conhecimento não trivial e aplicável. Entre os principais pontos de interesse quanto à busca de soluções, a redução de falhas na rede elétrica e a otimização de procedimentos executados para a operação da rede elétrica mostraram-se altamente significantes para a companhia. Isso se deu porque tal redução afetaria diretamente a qualidade da distribuição de energia aos consumidores e, por conseqüência, também os índices de qualidade definidos e regulados pela ANEEL.

Considerando que os especialistas já conheciam as classes de interesse dentro do domínio do problema – tipo de falhas, conseqüências, etc. –, restava-se caracterizar essas classes para se chegar a padrões no domínio de informação. Tais padrões teriam o objetivo de ajudar a prevenir interrupções de fornecimento e a melhor planejar os circuitos elétricos, evitando assim problemas envolvendo falhas já conhecidas. Por essa razão, entendeu-se que a tarefa de mineração de dados a ser aplicada sobre esse ambiente consistia de uma "classificação", a qual obteria regras de previsão sobre o comportamento da rede de distribuição.

Entre as técnicas para a obtenção de regras de classificação, a abordagem evolucionária apresentou-se como adequada ao se comparar às exigências do problema em relação às características intrínsecas que os algoritmos genéticos oferecem: possibilitar o processamento paralelo, considerar a interação entre atributos, permitir a parametrização dos seus métodos, adaptar-se às mudanças no ambiente populacional e gerar inovação nas soluções encontradas. Um algoritmo genético simples para classificação foi aplicado sobre dados de redes de baixa

tensão para testar o potencial da técnica ao trabalhar sobre o tipo de informação disponível. As regras de classificação obtidas foram consideradas válidas, e as necessidades que o escopo do problema possui foram mais bem evidenciadas após esse experimento.

Utilizando-se da existência na CELESC de uma base de dados que integra os seus sistemas transacionais e mantém informações históricas e não-voláteis (o DW), criou-se uma estrutura paralela (um EW) para a exploração e análise de dados orientada ao assunto de interrupções no fornecimento de energia, o *data mart* de operação. Da interação com os engenheiros especialistas foi feita uma amostragem de dados seguindo uma divisão condizente com a prática exercida na empresa e reunindo atributos, que, de acordo com eles, eram relevantes para o estudo de falhas da rede elétrica. Esses dados foram processados e transformados com o intuito de adequá-los aos objetivos da análise.

Um novo algoritmo genético, mais complexo e flexível do que o anterior, foi selecionado para aplicar a mineração: o sistema AGD – uma ferramenta genético-difusa que agrega as vantagens da Computação Evolucionária aos benefícios oferecidos pela Lógica Difusa no processamento e na apresentação de informações. A interface do protótipo, suas funcionalidades, estruturas de dados e o formato de entrada e saída de informações foram modificados para adaptarem-se ao tipo de processamento aqui exigido, bem como para adquirirem mais autonomia, confiabilidade e robustez na execução.

Após a definição dos conjuntos difusos para os atributos selecionados, a coleta das impressões gerais que os usuários possuem a respeito do problema e a definição dos parâmetros para o algoritmo genético, foram iniciados os testes sobre as amostras de dados. A partir dos três principais pontos de investigação levantados pelos especialistas, o AGD foi executado diversas vezes. As regras de classificação geradas foram extraídas, analisadas e selecionadas conforme o seu *fitness* (calculado sobre a qualidade da regra e o seu grau de interesse – ambos envolvendo os dados referentes à freqüência relativa, cobertura e taxa de acerto da regra).

As duas melhores regras geradas para os três assuntos foram aplicadas sobre a base de dados amostral. Considerou-se a hipótese de que seria possível reduzir pelo menos 10% dos problemas classificados pelas regras extraídas. Através do cálculo sobre a energia não-distribuída durante o intervalo causado por interrupções no fornecimento, os atributos dos

registros foram transformados em valores tangíveis, isto é, em termos monetários. O resultado foi apresentado também com relação aos índices de continuidade de distribuição afetados pela ocorrência de interrupções.

Os experimentos considerando apenas 10% dos casos abrangidos pelas regras de classificação aqui encontradas (Tabela 9.18) estimaram que a companhia elétrica deixou de arrecadar anualmente[14] a média total de 494.233,45 reais. Além da perda de receita, os mesmos experimentos calcularam que os custos sociais envolvidos com as interrupções classificadas foram em torno de 2.965.404,06 reais. Os valores para os índices de continuidade também são significativos: 2564,99 para DEC e 26,17 para FEC.

	Perda de Receita R$	Custo Social R$	DEC	FEC
Total	**494.233,45**	**2.965.404,06**	**2.564,99**	**26,17**

Tabela 9.19 – Total da perda de receita anual gerada por END e respectivos DEC e FEC causados

Através dos resultados obtidos no uso do AGD, inferências positivas podem ser apresentadas:

- Com relação às regras de classificação extraídas, sua validade, simplicidade de compreensão, utilidade prática, relevância no escopo do problema e o interesse que representa ao usuário, comprovam que a mineração de dados apresentada neste estudo atingiu a tarefa a que se propôs.

- Sobre algoritmos genéticos, a qualidade das regras encontradas também demonstra que esta é uma técnica não somente válida mas também eficaz para a extração de conhecimento no ambiente das redes de baixa tensão.

- Como contribuição deste trabalho em relação ao setor elétrico, ao utilizar técnicas de mineração de dados para a extração de conhecimento com valor estratégico e significativo, esta pesquisa pretende demonstrar a capacidade e a eficiência dessa

[14] Com base nos anos de 2004 e 2005.

abordagem tecnológica em conjunto com a valiosa experiência dos engenheiros eletricistas e outros especialistas nessa área.

– Quanto aos valores tangíveis encontrados, pretende-se que sirvam como exemplo do potencial da ferramenta utilizada, estimulando os especialistas a realizar muitas outras investigações, em maior nível e mais especificamente direcionadas.

A flexibilidade do AGD para aproveitar o conhecimento do especialista, dando liberdade a este para testar suas hipóteses, representa uma tentativa de reduzir a dependência e a distância entre o conhecedor do domínio de informação e o minerador de dados. Um papel não substitui o outro, mas ao disponibilizar meios tecnológicos ao especialista na área de negócio, difunde-se a aplicabilidade de soluções computacionais e abre caminho para novas demandas.

9.1 TRABALHOS FUTUROS

O uso de um sistema híbrido neste estudo é uma indicação de que agregar novos métodos aos já conhecidos pode trazer benefícios na busca por soluções. Por essa razão, adicionar novas funcionalidades ao AGD ou utilizar-se do conhecimento obtido aqui sobre o domínio das redes de baixa tensão para desenvolver uma nova técnica, capaz de melhor desempenhar a extração de regras de classificação, deve ser incentivado à medida que o uso da metodologia atual crie novas demandas.

Essa pesquisa delimitou o escopo da mineração sobre as redes de distribuição de baixa tensão, no entanto, outras áreas de informação do setor elétrico podem ser abordadas caso as devidas modificações sejam feitas em relação ao AGD. O *data warehouse* DW-Distribuição já possui *data marts* para outros assuntos envolvendo a rede elétrica, ou seja, é possível utilizar-se dessa estrutura existente na empresa para expandir o domínio a ser explorado. Fora do escopo de informações provido pela CELESC, essa pesquisa pode ser utilizada por outras companhias de distribuição de energia como referência à aplicação de mineração de dados, incentivando o desenvolvimento de inovações em técnicas e ferramentas para extração de conhecimento útil e relevante.

10 REFERÊNCIAS BIBLIOGRÁFICAS

ABERCROMBIE, M.; HICKMAN, C. J.; JOHNSON, M. L. **Diccionario de Biologia**. Barcelona: Labor S.A., 1970.

ADRIAANS, P. Z. ***Data Mining***. Harlow: Addison-Wesley, 1997.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining association rules between sets of items in large databases. In: ACM SIGMOD CONFERENCE, 1993, Washington, DC, USA. **Proceedings...** Washington, DC, USA, jun. 1993.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. In: 20th INTERNATIONAL CONFERENCE ON VLDB, 20., 1994, Chile. **Proceedins...** Chile, set. 1994. p. 487-499..

ANEEL. AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. Resolução Nº 318, de 6 de outubro de 1998.

_____. Resolução Nº 24, de 27 de janeiro de 2000.

_____. Resolução Nº 505, de 26 de novembro de 2001.

_____. Resolução Nº 456, de 29 de novembro de 2000.

AURÉLIO, Buarque de Holanda Ferreira. **Novo Dicionário Aurélio - Século XXI**. Nova Fronteira, 1999.

BAKER, J. E. **Adaptive selection methods for genetic algorithms**. In: GREFENSTETTE, J. J. (Ed.). Grefenstette ed., 1985.

BELLMAN. R. **Adaptive control processes:** A guided tour. Princeton: Princeton University Press, 1961.

BERRY, Michel J. A. ***Data Mining*** **techniques -** for marketing, sales, and customer support. New York: John Wiley & Sons, 1997.

BIGUS, Joseph P. ***Data Mining* with Neural networks**: Solving business problems from applications development to decision support. New York, NY: Computing McGraw-Hill, 1996.

BOOKER, L. B. Improving the performance of genetic algorithms in classifier systems. In: GREFENSTETTE, J. J. (Ed.). Genetic Algorithms and Their Applications. In: SECOND INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 2., 1996, Erlbaun. **Proceedings…** Erlbaun, 1996.

CABENA, P. et al. **Discovering *Data Mining*:** From Concept to Implementation. Upper Saddle River, NJ: Prentice Hall, 1998.

CELESC. **Definições e conceitos básicos utilizados na distribuição**. 1º CEDEN. Florianópolis, 198-.

_____. **Estudos sobre o tempo de mobilização das equipes de manutenção pesada de redes de distribuição frente a operação de distribuição**. Documentação interna. 01 dez. 2004.

_____. **Demonstrações Contábeis**. Exercícios findos em 31/12/2003 e 2002. Disponível em: <http://www.celesc.com.br/infofinanceiras/dc2003.pdf>. Acesso em: 22 nov. 2004.

_____. **Quem Somos**: A Empresa em Números. Agosto 2004. Disponível em: <http://www.celesc.com.br/quemsomos/numeros.php>. Acesso em: 22 nov. 2004.

CENTRAL VERMONT PUBLIC SERVICE. Glossary. Disponível em: <http://www.cvps.com/glossary.shtml>. Acesso em: 18 mar. 2005.

COLLARD, Martine; FRANCISCI, Dominique. **Evolutionary *Data Mining*:** an overview of Genetic-based Algorithms. França: IEEE, 2001.

CREDER, Hélio. **Instalações elétricas**. 11. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 1991.

DAVIS, L. D. **Handbook of Genetic Algorithms**. Van Nostrand Reinhold, 1991.

DE JONG, Kenneth A. **An Analisys of the Behavior of a Class of Genetic Adaptive Systems**. Ph.D. thesis. An Arbor: University of Michigan, 1975.

DUDA, Richard O. **Pattern classification and scene analysis**. EUA: John Wiley & Sons Inc, 1997.

FAYYAD, Usama M. *Data Mining* and knowledge discovery: making sense out of data. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC. EXPERT, 1996.

_____. *Data Mining* and Knowledge Discovery in Databases: Implications for Scientific Databases. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC. EXPERT, 1997.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From *Data Mining* to knowledge discovery: An overview. In: _____. **Advances in Knowledge Discovery and *Data Mining***. MIT, Cambridge, Massachusetts, and London, England: AAAI Press / The MIT Press, 1996a.

_____. Knowledge Discovery and *Data Mining*: Towards a Unifying Framework. In: SECOND INTERNATIONAL CONFERENCE ON KD & DM, 1996b, Portland, Oregon.

FOGEL, L. J.; OWENS, A. J.; WALSH, M. J. **Artificial Intelligence through Simulated Evolution**. Wiley, 1966.

FRAWLEY W.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. Knowledge Discovery in Databases: An Overview. **AI Magazine**, 1992.

FREIRE, Luciano Macedo. Análise de Custo/Benefício: Um Business Case Real aplicado à investimentos em automação de Subestações. **XV SNTPEE – Seminário Nacional de Produção e Transmissão de Energia Elétrica**. Outubro, 1999.

GARAI, Gautam; CHAUDHURI, B. B. **A novel genetic algorithm for automatic clustering**. Elsevier B. V., 2003.

GOLDBERG, David E. **Genetic algorithms in search, optimization, and machine learning**. EUA: Addison Wesley Longman, Inc., 1998.

GOLDBERG, David E.; RICHARDSON, J. Genetic Algorithms with sharing for multimodal function optimization. Genetic Algorithms and Their Applications. In: SECOND INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 2., 1997. **Proceedings…** Erlbaun, 1997.

GONÇALVES, Marcio E. **Uma Ferramenta de Extração de Dados para *data warehouse* Baseada em Agentes Distribuídos**. 2003. Dissertação (Mestrado em Computação), Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Santa Catarina. Florianópolis, 2003.

GONÇALVES, Alexandre L. **Utilização de Técnicas de Mineração de Dados em Bases de C&T**: Uma análise dos grupos de Pesquisa no Brasil. 2003. Dissertação (Mestrado em Engenharia de Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2003.

GORODETSKY, Vladimir; KARSAEYV, Oleg; SAMOILOV, Vladimir. **Software Tool For Agent-Based Distributed *Data Mining***. IEEE, 2003.

GUIMARÃES, Annelise Pegorini. **Estudo em Mineração de Dados aplicado a uma base de dados de materiais de estoque da indústria siderúrgica**. 2003. Trabalho de Conclusão (Tecnologia em Informática), Universidade Luterana do Brasil (ULBRA), Canoas, 2003.

HAND, D. **Construction and Assessment if Classification Rules**. Chichester: John Wiley & Sons, 1997.

HOLLAND, John. H. **Concerning efficient adptative systems**. Self-organizing systems. Washington DC: Spartan Books, 1962. p. 215-230.

_____. **Some pratical aspects of adptative systems theory**. Eletronic Information Handling. Washington, DC: Spartan Books, 1965. p. 209-217.

_____. **Universal Spaces:** A basis for studies of adaptation. Automata Theory. New York: Academic Press, 1966. p. 218-231.

_____. **Nonlinear Environments permitting efficient adaptation**. Computer and Information Sciences – II. New York: Academic Press, 1967. p. 147-164.

_____. **Hierarchical descriptions of universal spaces and adaptive systems**. Technical Report ORA Projects 01252 and 08226. Ann Arbor: University of Michigan, Department of Computer and Communication Sciences.

_____. **Goal-directed pattern recognition**. Methodologies of pattern recognition. New York: Academic Press, 1969. p. 287-296.

_____. Schemata and intrinsically parallel adaptation. In: NSF WORKSHOP ON LEARNING SYSTEM THEORY AND ITS APPLICATIONS, 1973, Gainesville. **Proceedings…** Gainesville: University of Florida, 1973. p. 43-46.

_____. **Adaptation in natural and artificial systems**. Ann Arbor: University of Michigan Press, 1975.

_____. Escaping brittleness. In: MICHALSKI, R. S.; CARBONELL, J.; MITCHEL, T. (Ed.). **Machine Learning**: An artificial intelligence approach. San Mateo, CA: Morgan Kaufmann, 1986. v. 2.

HUANG, Yin-Fu; WU, Chiech-Ming. Mining Generalized Association Rules Usin Puning Techniques. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., 2002.

INMON, William H. **Como Construir o *data warehouse***. EUA: Wiley Computer Publishing, 1997.

INMON, William. H; TERDEMAN, Robert H.; IMHOFF, Claudia. **Data Warehousing**: Como transformar informações em oportunidades de negócios. Tradução de Melissa Kassner. São Paulo: Berkeley Brasil, 2001.

JANIKOW, C. Z.; MICHALEWICZ, Z. **An experimental comparison of binary and floating representation in genetic algorithms**. In: BELEW, R. K.; BOOKER, B. (Ed.). 1991.

KIMBALL, Ralph. **The *data warehouse* toolkit**: the complete guide to dimensional modeling. EUA: Wiley Computer Publishing, 2002.

KING, R. D. et al. **STATLOG:** Comparison of Classification algorithms on large real-world problems. Applied Art. Int 9(3), May/June 1995. p. 289-333.

KOZA, J. R. **Genetic Programming**: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.

LIU, Bing; HSU, Wynne. Post-analysis of learned rules. In: AAAI-96, p. 828-834.

LUCAS, Anelise de Macedo. **Utilização de Técnicas de Mineração de Dados considerando os aspectos temporais**. 2002. Dissertação (Mestrado em Ciência da Computação), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

MATHEUS, Christopher J.; CHAN, Philip K.; PIATETSKY-SHAPIRO, Gregory. Systems for Knowledge Discovery in Databases. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., 1993.

MITCHELL, Melanie. **An introduction to genetic algorithms**. MIT, 1996.

MUSEUM OF SCIENCE BOSTON BOSTON. **Glossary of technical terms**. Disponível em: <http://www.mos.org/sln/toe/glossary.html>. Acesso em: 18 mar. 2005.

PACITTI, Tércio; ATKINSON, Cyril P. **Programação e métodos computacionais**. 2. ed. Rio de Janeiro: Livros Técnicos e Científicos Editora S/A, 1977. v. 2.

PIATETSKY-SHAPIRO, G.; MATHEUS, C. The Interestingness of Deviations. In: PROCEEDINGS OF KDD-94 WORKSHOP, AAAI Press, 1994.

PIATETSKY-SHAPIRO, Gregory. Knowledge Stream Partners. The Data-Mining Industry Coming of Age. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC. EXPERT OPINION, 2000.

PUBLIC POWER COUNCIL. **Glossary of terms**. Disponível em: <http://www.ppcpdx.org/Sidebar/Glossary2.htm#Terms%20Letter%20S>. Acesso em: 18 mar. 2005.

RICHARDS, G.; RAYWARD-SMITH, V. J. Discovery of Association Rules in Tabular Data. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., 2001.

ROMÃO, Wesley et al. Algoritmos genéticos e conjuntos difusos aplicados ao controle de um processo térmico. **Revista Tecnológica**, n. 8, p. 7-21, 1999b.

ROMÃO, Wesley. **Descoberta de Conhecimento Relevante em Banco de Dados sobre Ciência e Tecnologia**. 2002. Tese (Doutorado em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis, 2002.

ROMAO, Wesley; FREITAS, Alex A.; PACHECO, Roberto C. S. Uma revisão de abordagens genético-difusas para descoberta de conhecimento em banco de dados. **Proceedings...** 2002.

SALVADOR, Otávio. **Introdução a Algoritmos Genéticos**. Rio Grande do Sul: Universidade Católica de Pelotas.

SCHALKOFF, Robert J. **Pattern Recognition:** statistical, structural and neural approaches. EUA: John Wiley & Sons, Inc., 1992.

SCHNEIDER, André M. **Algoritmo Adaptativo Genético para Acompanhamento da Trajetória de Alvos Móveis**. 1998. Dissertação (Mestrado) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1998.

SCHREIBER, August T. et al. **Knowledge Engineering and Management: the CommonKADS methodology**. MIT Press, 2000. Chapter 1,

SCHUSTER, Assaf; WOLFF, Ran; TROCK, Dan. A High-Performance Distributed Algorithm for Mining Association Rules. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., 2003.

SINGH, Y. P.; ARABY, Norhana A. R. Evolutionary Approach to *Data Mining*. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC., 2000.

SUPER INTERESSANTE, mar. 2004.

TODESCO, José L. et. al. Uma Plataforma de Gestão de Redes de Distribuição de Baixa Tensão. In: XXIV ENEGEP, 2004a, Florianópolis.

TODESCO, José L. et al. Gestão de Distribuição Secundária de Energia Elétrica utilizando um Sistema Especialista. In: XXIV ENEGEP, 2004b, Florianópolis.

TODESCO, José L. et al. Previsão de Demanda de Energia usando Famílias de Circuitos e Rede Neural Artificial. In: XXIV ENEGEP, 2004c, Florianópolis.

TSOUKALAS, L. H.; UHRIG, R. E. **Fuzzy and Neural Approaches in Engineering**. New York: John Wiley and Sons, 1997. Chapter 5.

TWO CROWS CORPORATION. **Introduction to *Data Mining* and Knowledge Discovery**. 3. ed. 1999.

WILLIAMS, Graham et al. **A Comparative Study of RNN for Outlier Detection in *Data Mining***. IEEE, 2002.

WITTEN, Ian H.; FRANK, Eibe. ***Data Mining***: practical machine learning tools and techniques with Java implementations. EUA: Morgan Kaufmann Publishers, 2000.

WRIGHT, A. H. **Genetic Algorithms for real parameter optimization**. In: RAWLINS, G. (Ed.). 1991.

XIONG, N.; LITZ, L. Generating Linguistic Fuzzy Rules for Pattern Classification with Genetic Algorithms. In: PKDD-99, 1999. p. 574-579.

YEPES, Igor. **Projeto ISIS – Temas Inteligentes:** Uma incursão aos Algoritmos Genéticos. Disponível em: <http://www.geocities.com/igoryepes/visualizar2.htm#operador>. Acesso em: 30 nov. 2004.

ZADEH, L. A. **Fuzzy Sets. Information and Control.** 1965**.** p. 338-353.

11 APÊNDICE

11.1 CAUSAS DE INTERRUPÇÃO DE ENERGIA ELÉTRICA

A lista das causas naturais e não-naturais (consideradas como previsíveis), ordenada por número de ocorrências existentes na base de dados, é apresentada na Tabela 11.20. É importante observar o critério utilizado para separação das causas, já que há algumas descritas como indefinidas. Por serem causas avaliadas pelo técnico que efetuou a checagem da reclamação do consumidor ou executou a correção do problema, a descrição das causas vai desde definições generalizadas até detalhes técnicos encontrados. No entanto, existe um grande número de causas não resolvidas, isto é, situações em que o técnico não pôde estimar o motivo da falha. Isso ocorreu porque muitas vezes a causa do problema não está mais presente quando o técnico chega ao local.

Um exemplo simples ocorre quando, devido à sobrecarga de tensão em um cabo, a alta temperatura causa uma expansão no material do cabo, fazendo com que ele toque outro cabo próximo e gere curto-circuito. Os dispositivos de segurança no transformador automaticamente interrompem a alimentação elétrica naquele cabo. Sem eletricidade, a temperatura diminui e o material volta à sua condição normal, afastando-se do cabo próximo. Quando o técnico chega ao local, ele não vê mais a fonte do problema. Esse é apenas um dos muitos exemplos ao qual se juntam as falhas geradas por motivos climáticos (vendaval), do meio ambiente (animal e vegetal) e do meio urbano (pipa presa no fio e abalroamento).

Número de Ocorrências	Ocorrências %	Nome da causa	Pode ser prevenida?
37097	14,8	A INVESTIGAR	Não
34861	13,91	DESCARGA ATMOSFERICA - FENÔMENO NATURAL	Não
23345	9,31	MEIO AMBIENTE ANIMAL	Não
18695	7,46	VEGETAÇÃO NA REDE - MEIO AMBIENTE	Sim
15398	6,14	PROG. - MANUTENÇÃO PREVENTIVA	Sim
13984	5,58	TERCEIROS - PIPA, BOLA, ... (ESPECIFICAR)	Não
13761	5,49	PROG. - ALTERAÇÃO PARA MELHORIA	Sim
13021	5,19	PROG. - ALTERAÇÃO PARA AMPLIAÇÃO	Sim
10289	4,1	PROG. - MANUTENÇÃO CORRETIVA - EMERGÊNCIA	Sim
9803	3,91	VENDAVAL	Não
6808	2,72	ABALROAMENTO	Não
4334	1,73	DEFEITO EM PARA-RAIO	Sim
4173	1,66	FALHA EM CHAVE FUSÍVEL (FROUXA, MA CONEXAO ,OXID)	Sim
4065	1,62	DEFEITO EM ISOLADOR - TRINCADO, QUEBRADO	Sim
3888	1,55	DEFEITO EM CONDUTOR - EXECESSO EMENDAS, VELHO	Sim
3668	1,46	CONDUTOR DESREGULADO	Sim
3569	1,42	OUTROS ¿ OCORRÊNCIAS EM REDE (ESPECIFICAR)	Não
3482	1,39	FALHA EM ELO (INADEQUADO, FADIGA,DESREGULADO,...)	Sim
2583	1,03	SOBRECARGA NO TRANSFORMADOR	Sim
2476	0,99	CQDE - TAP INADEQUADO	Sim
2363	0,94	DEFEITO NO TRANSFORMADOR (INTERNO, FERRUGEM, BUCHA	Sim
2354	0,94	MÁ CONEXÃO NA REDE SECUNDÁRIA	Sim
2238	0,89	POSTE AVARIADO, CAIDO, PODRE, OU FORA DE PRUMO	Sim
2207	0,88	JAMPER OU FLY-TAP PARTIDO	Sim
1791	0,71	OUTROS COMPONENTES	Não
1394	0,56	ACIDENTAIS - TRANSMISSÃO (>= 3 MINUTOS)	Não
1243	0,5	MÁ CONEXÃO NOS BORNES DO TRANSFORMADOR	Sim
1180	0,47	CRUZETA : PODRE, QUEIMADA OU QUEBRADA	Sim
692	0,28	ISOLADOR - MEIO AMBIENTE CLIMA (SALITRE, NEVE)	Sim
591	0,24	ROMPIMENTO DE CONDUTOR DEVIDO AO FRIO	Sim
545	0,22	ACIDENTAIS - SUPRIMENTO (TRANSMISSÃO)	Não
523	0,21	MÁ CONEXÃO NA REDE PRIMÁRIA	Sim

Número de Ocorrências	Ocorrências %	Nome da causa	Pode ser prevenida?
437	0,17	PROG. - TRANSMISSÃO (>= 3 MINUTOS)	Sim
382	0,15	FALHA EM CHAVE-FACA : FROUXA, MÁ CONEXAO ,OXIDADA	Sim
374	0,15	EXECUTAR MANOBRAS	Sim
302	0,12	ISOLADOR SUJO (POLUICAO OU POEIRA)	Sim
248	0,1	MÁ CONEXÃO COM CONECTOR NO RAMAL DE LIGAÇÃO	Sim
213	0,08	MÁ CONEXÃO EM CHAVE	Sim
211	0,08	PROG. - SUPRIMENTO (TRANSMISSÃO)	Sim
182	0,07	CURTO CIRCUITO NA TUBULAÇÃO - RAMAL DE ENTRADA	Sim
169	0,07	CQDE - SOBRECARGA NOS CONDUTORES	Sim
161	0,06	DEFEITO NO REGULADOR DE TENSÃO	Sim
158	0,06	RAMAL DE LIGAÇÃO PARTIDO	Sim
158	0,06	PODA DE ÁRVORE NA BT OU RAMAL DE LIGAÇÃO	Sim
136	0,05	PROG. - AUMENTO DA POTÊNCIA	Sim
133	0,05	RAMAL DE LIGAÇÃO EM CURTO CIRCUITO	Sim
106	0,04	MÁ CONEXÃO FIO A FIO NO RAMAL DE LIGAÇÃO	Sim
103	0,04	ILUMINAÇÃO PÚBLICA	Sim
83	0,03	DISJ QUADRO MEDICAO C/ DEFEITO - MÁ CONEX, QUERADO	Sim
78	0,03	CQDE - CIRCUITO DESBALANCEADO	Sim
64	0,03	CQDE - CIRCUITO EXTENSO (LONGE DO CONSUMIDOR)	Sim
54	0,02	OUTROS - RAMAL DE LIGAÇÃO (ESPECIFICAR)	Não
52	0,02	PROG. - DESMONTE DE OBRAS	Sim
49	0,02	CQDE - CABO SAÍDA TRAFO INADEQUADO (CABO VPP)	Sim
47	0,02	PROG. - SUBSTITUIÇÃO POR POTÊNCIA MENOR	Sim
45	0,02	POSTE INTERMEDIÁRIO CAÍDO OU PODRE	Sim
38	0,02	CQDE-TRANSFORMADOR DESCENTRALIZADO DO CIRCUITO	Sim
35	0,01	OUTROS NO QUADRO DE MEDIÇÃO (ESPECIFICAR)	Não
31	0,01	OUTROS DO RAMAL DE ENTRADA (ESPECIFICAR)	Não
30	0,01	CQDE - CIRCUITO EM MAU ESTADO (PIPA,RECOZ, ETC.)	Não
29	0,01	DEFEITO NOS BORNES DO MEDIDOR (MÁ CONEXÃO, QUEBRAD	Sim
22	0,01	QDDE - DEFEITO NO CABO DE SAÍDA DO TRAFO (CABOVPP)	Sim
22	0,01	CQDE - CAPACITOR AT DESLIGADO	Sim
20	0,01	DEFEITO INTERNO NO MEDIDOR	Sim

Número de Ocorrências	Ocorrências %	Nome da causa	Pode ser prevenida?
19	0,01	CQDE - ATERRAMENTO INADEQUADO	Sim
18	0,01	RAMAL DE LIGAÇÃO TRANÇADO/ DESREGULADO	Sim
18	0,01	PROG. - TRANSFORMADOR SEM CARGA	Sim
13	0,01	PROG. - DIVISÃO DE CIRCUITO	Sim
9	0	PROG. - MANUT PREVENTIVA POR CORROSÃO FUNDO TANQUE	Sim
7	0	CQDE - CONS PROVOCANDO PERT SISTEMA-MOTOR,BATE ES	Sim
6	0	PROG. - MANUT. PREVENTIVA POR CORROSÃO EM RADIADOR	Sim
4	0	BRAQUETE SOLTA	Sim
2	0	PROG. - MANUT. PREVENTIVA POR CORROSÃO NA TAMPA	Sim
1	0	CQDE - RAMAL DE ENTRADA EXTENSO (APÓS MEDIÇÃO)	Sim
1	0	FUSÍVEL QUADRO MEDIÇÃO QUEIMADO	Sim

Tabela 11.20 - Lista das causas de interrupção elétrica

11.2 CONSULTAS SQL PARA OS CÁLCULOS SOBRE A ENERGIA NÃO-DISTRIBUÍDA

Para realizar os cálculos relativos à END de forma automática pelo banco de dados, foram adicionadas à consulta constantes numéricas conforme o cálculo desejado e seus aspectos:

- END: valor relativo à tarifa elétrica vigente no Estado (R$ 0,40);
- Custo Social: valor referente a R$ 2,40 – conforme estimado por Freire (1999);
- Porcentagem de registros corrigidos: valor de 0.1 para obter apenas 10% dos registros;
- Unidade da END: valor indicando 60 minutos para a transformação da unidade do cálculo em quilowatt/hora.

A seguir têm-se as consultas SQL aplicadas à base de dados para obtenção dos cálculos necessários para as três melhores regras selecionadas pelo analista (seção 8.2). As constantes utilizadas para o cálculo estão em vermelho e em negrito.

a) Interrupções com relação ao período do dia

Regra 1

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND HORA_INI >= 4
      AND IN_DEC >= 0.8
       AND ID_TN_NOMINAL_AL = 1)
```

Regra 2

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND HORA_INI >= 4
       AND MTRMT_SE IS NULL
      AND IN_DEC BETWEEN 0.1 AND 0.79)
```

b) Sazonalidade das causas

Regra 1

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE ID_MES BETWEEN 8 AND 12
       AND ID_DIA_SEMANA >= 5
       AND PO_TOTAL <= 245650
       AND DS_CAUSA = 'VEGETAÇÃO NA REDE - MEIO AMBIENTE')
```

Regra 2

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE
      ID_MES BETWEEN 5 AND 7
       AND ID_DIA_SEMANA BETWEEN 5 AND 7
       AND DS_CAUSA = 'SOBRECARGA NO TRANSFORMADOR')
```

c) Potência interrompida por manutenção programada

Regra 1

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND DS_CAUSA = 'PROG. - ALTERAÇÃO PARA MELHORIA'
           AND PO_INTERROMPIDA >= 3333)
```

Regra 2

```
SELECT (SUM(VALOR) * 0.4) FROM (
SELECT ROUND(((PO_INTERROMPIDA * (QT_MN_INTERRUPCAO))/60)*0.1, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND DS_CAUSA = 'PROG. - MANUTENÇÃO PREVENTIVA'
       AND QT_CONSUMIDOR_INTERRUPCAO <= 1353
       AND QT_MN_INTERRUPCAO <= 591.18
        AND PO_INTERROMPIDA <= 1650)
```

Para o cálculo do custo social, apenas substituiu-se a constante “0.4” pela constante “2.4”. Quanto ao cálculo do DEC e FEC, somente a constante relativa a 10% de registros foi mantida na consulta, não sendo mais necessário envolver os atributos de potência e minutos interrompidos, como nos exemplos a seguir referentes às consultas de DEC e FEC (respectivamente) para a Regra 2 do assunto “Potência interrompida por manutenção programada”:

DEC

```
SELECT SUM(VALOR*0.1) FROM (
SELECT ROUND(IN_DEC, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND DS_CAUSA = 'PROG. - MANUTENÇÃO PREVENTIVA'
       AND QT_CONSUMIDOR_INTERRUPCAO <= 1353
       AND QT_MN_INTERRUPCAO <= 591.18
       AND PO_INTERROMPIDA <= 1650)
```

FEC

```
SELECT SUM(VALOR*0.1) FROM (
SELECT ROUND(IN_FEC, 2) AS VALOR
FROM DM_REGISTRO_ANALISE
WHERE PERIODO_INI = 'Manhã'
       AND DS_CAUSA = 'PROG. - MANUTENÇÃO PREVENTIVA'
       AND QT_CONSUMIDOR_INTERRUPCAO <= 1353
       AND QT_MN_INTERRUPCAO <= 591.18
       AND PO_INTERROMPIDA <= 1650)
```