DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE EVENTOS DE DESLIGAMENTOS DE EMPRESAS DE DISTRIBUIÇÃO

Alex B. Tronchoni* **Carlos O. Pretto***

Mauro A. da Rosa* **Flávio A. Becon Lemos***

*Grupo de Sistemas de Energia Elétrica - GSEE
Faculdade de Engenharia - PUCRS
Av. Ipiranga 6681, Prédio 30, Partenon
CEP 90619-900 - Porto Alegre RS

RESUMO

Este artigo explora a técnica de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases – KDD*) com o objetivo de qualificar a informação recolhida durante os trabalhos de recomposição de sistemas de distribuição por equipes de eletricistas. Esta qualificação possibilita a utilização de técnicas de Inteligência Artificial (*IA*) para apoiar decisões de investimentos em planejamento, operação e manutenção de sistemas de distribuição. Com o objetivo de ilustrar a importância dessa qualificação, este artigo apresenta, adicionalmente, a utilização dos resultados da aplicação de *KDD* para o treinamento de uma Rede Bayseana (*RB*). A meta principal da *RB* é auxiliar no diagnóstico de desempenho das redes elétricas, promovendo uma identificação indireta de causas de desligamentos forçados. A análise dos dados coletados durante uma interrupção forçada de energia elétrica indica que o principal objetivo dos eletricistas em campo é a rápida recomposição da rede e, por muitas vezes, as causas que cercam os eventos que originaram as interrupções possuem um alto nível de subjetividade e incerteza, impossibilitando a sua identificação direta. Para ilustrar essa metodologia é apresentado um caso com 570.000 eventos, ao qual o *KDD* proporciona um novo ambiente – com um número significativo de dados – mais apropriado para o treinamento e validação da *RB* para identificação de causas de desligamentos não programados.

PALAVRAS-CHAVE: Descoberta de conhecimento em base de dados, Mineração de dados, Banco de dados, Identificação da causa de desligamentos não programados, Redes Bayesianas.

ABSTRACT

Knowledge Discovery In Databases of Forced Outage of Distribution Utilities

This paper tackles the Knowledge Discovery in Databases (KDD) technique in order to qualify the information collected by the field crews during the restoration process of the distribution network. This improvement of the information enables the utilization of Artificial Intelligence (AI) techniques to support decision making in investments planning, operation and maintenance of distribution systems. This qualification allows the use of KDD results to train a Bayesian Network (BN). The aim of the proposed BN is to support the performance diagnostic of electrical networks, promoting an indirect identification of forced outage causes. The analysis of the collected data during a forced outage indicates that the main purpose of field crews is a quick restoration of the network and, several times, the causes surrounding the outage events have a high level of subjectivity and

Artigo submetido em 07/01/2008 (Id.: 00846)
Revisado em 01/08/2008, 24/11/2008, 12/02/2009, 17/05/2009
Aceito sob recomendação do Editor Associado Prof. Julio Cesar Stacchini Souza

uncertainty, turning impossible its direct identification. To illustrate the methodology, is presented a case with 570.000 events, where KDD provides a new environment – with a significant amount of data – more suitable to train and validate a BN to identification of forced outage causes.

KEYWORDS: Knowledge discovery in databases, Data mining, Database, Forced outage cause identification, Bayesian Network.

1 INTRODUÇÃO

A necessidade de armazenar e tratar dados de diversas fontes é um desafio cada vez maior no setor elétrico (Wehenkel, 1998). A competitividade imposta nos anos 90, através das privatizações das empresas de distribuição de energia, fez crescer em importância áreas como Tecnologia da Informação (TI), Logística (Materiais e Equipes), Mercados (Compra e Venda de Energia) entre outras, ampliando a gama de controle necessário às informações provenientes dessas áreas (Billinton *et alii*, 1995). O volume de informações relacionadas aos aspectos regulatórios, controle e cadastro de equipamentos e redes, monitoração de dados da operação, automação e controle das redes de distribuição e de muitas outras fontes, fez aumentar significativamente a dimensão das bases de dados das empresas de energia. A utilização adequada dessas informações se tornou fundamental na busca pelo aumento da eficiência e da produtividade deste segmento empresarial (Wehenkel, 1998). Entretanto, essa nova condição torna difícil e trabalhosa a tarefa de manipulação e utilização adequada das informações contidas nessas bases de dados. Em muitos casos, as condições de obtenção dessas informações (coleta de dados) são deficientes e precárias, comprometendo a qualidade do dado armazenado e posteriormente a sua utilização concreta. É possível dizer que uma das dificuldades na área de desenvolvimento de softwares, voltados às empresas de distribuição de energia, tem sido encontrar informações qualificadas que possam servir de dados de entrada para ferramentas de apoio à decisão (Pretto *et alii*, 2006). Uma destas situações é a identificação correta das causas de um desligamento não programado, a qual está diretamente relacionada a uma gama de fatores que podem ou não ser perceptível ao eletricista (coletor) no local do evento. Entre estes fatores, destaca-se a subjetividade da informação adquirida no local do evento, as condições atmosféricas que cercam o evento, o histórico de recorrência do evento, a identificação correta das atividades e dos fatos que envolvem a rede e que poderiam indicar uma causa de interrupção de energia. Muitas vezes a atribuição de uma causa de desligamento está associada a fatos locais aparentes, mas que na realidade podem ser a conseqüência ou causa secundária para o registro do evento, e terem sido desencadeados por uma falta primária não perceptível na inspeção local durante o restabelecimento de energia. Outras vezes, por não haver uma causa aparente, é comum apontar para uma falta transitória, como motivo da interrupção. Um exemplo típico é o desligamento por curto-circuito. Ao restabelecer o fornecimento de energia após detectar o curto-circuito que originou a atuação da proteção, o eletricista (coletor) pode anotar na sua ficha de serviço como causa da interrupção a atuação da proteção e/ou o curto-circuito. Na realidade, o curto-circuito e a atuação da proteção são conseqüências de algum evento anterior que ocorreu na rede, tais como condutores sem tração adequada, isoladores danificados, e muitos outros fatores que podem interromper o fornecimento de energia. A correta identificação da causa raiz, ou causa primária, é tão ou mais importante que o próprio restabelecimento, pois muitas vezes permite identificar a recorrência de um conjunto de causas que podem estar originando a degradação acelerada da rede e, por conseqüência, dos indicadores de Duração Equivalente de Interrupção por Unidade Consumidora (DEC) e Freqüência Equivalente de Interrupção por Unidade Consumidora (FEC), orientando as ações de planejamento, operação e manutenção de forma direta e econômica e, consequentemente, a melhoria dos indicadores de continuidade do fornecimento de energia.

Geralmente, em empresas de distribuição de energia, a tarefa de identificação e coleta de causas de desligamentos não programados (interrupções fortuitas) é realizada pelos eletricistas. Posteriormente, essas causas observadas são armazenadas em uma base de dados em estado bruto, sem qualquer tipo de filtragem ou tratamento, sendo apenas afetadas por uma validação e consolidação manual, atribuída ao responsável pela passagem das informações escritas nas fichas de serviço, para o formato digital. Notadamente, o percurso da informação pode, ocasionalmente, apresentar falhas no preenchimento, na interpretação ou na leitura da ficha de serviço, bem como na própria passagem para o formato digital. A conseqüência dessa falha é percebida posteriormente, com a tentativa de utilização dessas informações no estado coletado, as quais geralmente apresentam inconsistências, dados dispersos, incompletos e na grande maioria conflitantes. A importância dessas observações pode ser melhor compreendida quando confrontadas com os números de eventos registrados pelas empresas. Em termos médios, uma empresa de distribuição, com redes suficientes para o fornecimento de energia de um milhão de consumidores, registra um número médio de eventos não programados de cerca de 200.000 ao ano. Dessa forma, é simples notar que a falta de uma metodologia de organização e classificação, inviabiliza uma utilização ou análise adequada dessa grande quantidade de dados coletados em campo, fazendo com que sejam perdidas importantes correlações entre informações com a mesma origem primária. Por outro lado, a utilização de métodos de avaliação automáticos requer uma organização sistemática das diversas fontes de informações, de tal forma que possam

ser realizadas correlações entre os dados e criados conjuntos para treinamento e validação desses métodos, tais como Redes Neurais Artificiais (RNA) (Haykin, 2001), Redes Bayesianas (RB) (Norvig e Russel, 2003) e Árvores de Decisão (Norvig e Russel, 2003).

Com o objetivo de auxiliar na criação de uma metodologia capaz de classificar e identificar causas primárias de desligamentos não programados foi necessário realizar um processo de identificação, depuração, transformação, seleção e validação de informações em bases de dados. Naturalmente, a necessidade conduziu à aplicação de uma técnica de extração de conhecimento em banco de dados, também conhecida como *Knowledge Discovery in Databases* (*KDD*) (Fayyad *et alii*, 1996).

Este artigo explora a técnica de *KDD* para executar uma organização de dados coletados durante o restabelecimento de redes aéreas de distribuição de energia. A partir de uma base de dados em estado bruto, gera uma nova base de dados dedicada à identificação automática de causas primárias de desligamentos não programados. Para ilustrar essa metodologia é apresentado um caso com 570.000 eventos, ao qual a metodologia proporciona um novo conjunto de dados específico para o treinamento e validação da ferramenta automática de identificação de causas de desligamentos não programados.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS - *KDD*

Devido a sua natureza em dimensão, complexidade e diversidade de dados, a área de sistemas de energia possui características adequadas para utilização de sistemas baseados em inteligência computacional para detecção e localização de faltas em redes de distribuição, tais como sistemas especialistas (Hsu *et alii*, 1990), Redes Bayesianas (Chien *et alii*, 2002), *Rough Set* (Peng *et alii*, 2004), Redes Neurais Artificiais (Chow *et alii*, 1993) (Xu e Chow, 2006), Mapas Auto-Organizáveis (Costa e Andrade Netto, 2007), entre outros. Entretanto, a maioria desses métodos não está interessada no diagnóstico da falta, mas no apoio ao processo de identificação e melhoria na segurança e nos tempos de restabelecimento, através da orientação das equipes sobre a localização dos defeitos. Evidentemente, esses algoritmos exploram as dificuldades inerentes à identificação visual das faltas, nomeadamente para redes subterrâneas, porém, sem o objetivo principal de identificar as causas que levaram à interrupção. Dessa forma, o objetivo de qualificar a informação contida em bases de dados relativas a interrupções é sem dúvida uma necessidade. Recentemente, uma maior atenção tem sido direcionada neste sentido, como demonstra a referência (Xu *et alii*, 2007). A técnica de *KDD* é um processo de identificação de padrões de dados válidos, não-triviais, potencialmente úteis e compreensíveis (Fayyad *et alii*, 1996). O processo de

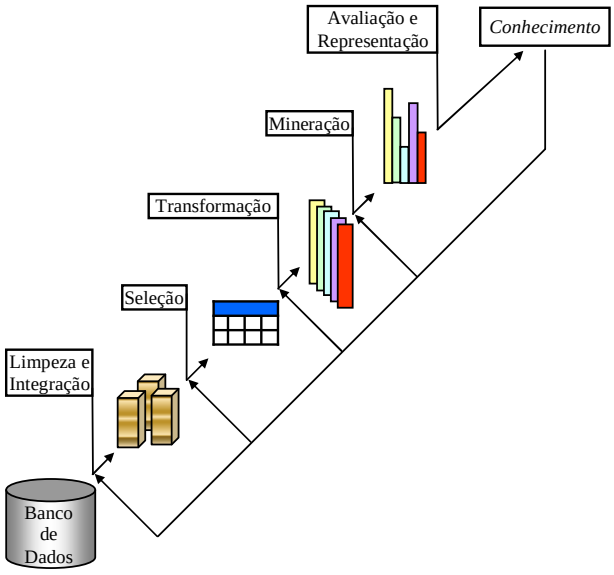

Figura 1: Fluxo do Processo de KDD

KDD é apresentado na Figura 1, e consiste em uma sequência iterativa dos seguintes passos (Han e Kamber, 2000):

1. **Limpeza dos dados**: para remover ruídos e dados irrelevantes;
2. **Integração dos dados**: onde fontes de dados múltiplos podem ser combinadas;
3. **Seleção dos dados**: onde dados relevantes para a análise são recuperados do banco de dados;
4. **Transformação dos dados**: onde os dados são transformados ou consolidados no formato apropriado para mineração;
5. **Mineração de dados**: é um processo onde métodos inteligentes são utilizados a fim de extrair padrões de dados;
6. **Avaliação e representação do conhecimento**: onde técnicas de visualização e representação de conhecimento são utilizadas para apresentar o conhecimento extraído para o usuário.

Alguns autores consideram *KDD* e Mineração de Dados como processos distintos (Fayyad *et alii*, 1996). Entretanto, em algumas bibliografias, o termo mineração de dados (*Data Mining*) tornou-se mais popular que o *KDD* e é utilizado quando se refere ao processo de identificação de padrões a partir de grandes quantidades de dados armazenados em bancos de dados ou outro tipo de banco de armazenamento (Han e Kamber, 2000).

3 BASE DE DADOS DE DESLIGAMENTOS

A criação e a manutenção de uma base de dados sobre o histórico das intervenções na rede de distribuição é uma importante tarefa nas empresas de distribuição energia. As informações dessa base de dados são geralmente coletadas pelas equipes de eletricistas, e o histórico das atuações das equipes de emergência ou de manutenção devem ser corretamente armazenados. É comum encontrar nas empresas um padrão de classificação de causas de interrupção, o qual promove suporte ao pessoal de campo no apontamento das possíveis causas de desligamentos e facilita a transformação do dado escrito para o formato digital. A adoção de processos que proporcionam uma forma organizada de armazenar o histórico desses eventos é importante para orientar ações de planejamento, operação e manutenção.

Na literatura existem algumas classificações de causas, como a descrita em Oprisan *et alii* (1991), onde o autor classificou as causas em função de sua natureza, a qual resultou nas seguintes 10 categorias:

- Outras – Interrupções sem uma causa ou razão aparente que possa ter contribuído para o desligamento;
- Interrupção Programada – Interrupção de fornecimento devido ao desligamento programado para manutenção preventiva ou construção;
- Perda de Suprimento – Interrupção de fornecimento devido a problemas no sistema supridor causados pela diminuição da freqüência por aumento de carga, tensão fora dos limites aceitáveis de operação, transitório no sistema de transmissão entre outros;
- Contato com Árvores – Interrupção originada pelo contato de árvores no circuito elétrico;
- Descarga Atmosférica – Interrupção de fornecimento devido a descargas atmosféricas;
- Equipamento Defeituoso – Interrupção causada por falha de equipamentos devido ao tempo de uso, manutenção incorreta ou falhas eminentes;
- Clima Adverso – Interrupção de fornecimento de energia causada por chuva, gelo, tempestades, neve, vento, temperaturas extremas, ou outra condição climática crítica;
- Ambiente Adverso – Interrupção causada por exposição dos equipamentos a condições anormais, como maresia, contaminação industrial, umidade, corrosão, vibração, fogo ou enchente;
- Elemento Humano – Interrupção de fornecimento devido à interferência de funcionários da empresa, uso incorreto de equipamentos, instalação ou construção incorreta, erro de configuração de proteção ou erro de manobras de chaveamento;
- Interferência Externa – Interrupção de energia devido a fatores fora do controle da concessionária, como pássaros e animais diversos, veículos, escavações, vandalismo, sabotagem e objetos estranhos.

Em Brown (2002), o autor cita as causas mais comuns de desligamentos em redes (aéreas e subterrâneas) de distribuição de energia, fazendo um comentário sobre essas causas e seus impactos no desempenho da rede. A próxima seção apresenta a classificação de causas proposta e utilizada neste artigo.

3.1 Classificação Proposta de Causa de Desligamentos

Como comentado anteriormente, é de fundamental importância a definição e criação de uma estrutura que classifique as causas que provocam desligamentos em sistemas de distribuição. Neste trabalho foram analisadas e tomadas como base para classificação as referências Oprisan *et alii* (1991), Brown (2002) e Pretto (2005b) onde as causas são agrupadas de acordo com a natureza do evento e sua contribuição para a ocorrência da falta de energia. Uma nova classificação de causas baseada na seção 3 é proposta neste artigo, onde as causas foram reestruturadas para uma melhor adequação ao trabalho. As classes de causas de desligamentos definidas para este trabalho são as seguintes:

1. *Falha no Componente* Interrupção de fornecimento causada por falha do componente (transformador, fusível, capacitor, etc) devido aos mais diversos fatores;
2. *Sobrecarga* Interrupção causada por sobrecarga do sistema. Em dias de muito consumo, pode ocorrer a atuação de uma chave protetora em função do carregamento das redes;
3. *Clima Adverso* Interrupção de fornecimento de energia causada por chuva, granizo, tempestades, ventos, temperaturas extremas, ou outras condições climáticas adversas que possam influenciar num desligamento;
4. *Descarga Atmosférica* Interrupção causada por uma descarga atmosférica na rede elétrica;
5. *Interferência do Meio Ambiente* Interrupção ocasionada por intervenção do meio ambiente no entorno da ocorrência do desligamento, tais como objetos presos na rede elétrica, animais, entre outros;

6. *Vegetal* Interrupção causada por influência específica de árvores, galhos e arbustos perto da rede elétrica;

7. *Interferência Humana* Quando ocorrem desligamentos ocasionados por vandalismo, furto ou outra interferência humana é atribuído esta causa. Interrupções ocasionadas por empresas que trabalham perto da rede, obras, empresas de telefonia entre outras, também são atribuídas a este item;

8. *Acidente* Interrupções ocasionadas por acidentes envolvendo veículos;

9. *Incêndio/Queimada* Interrupções ocasionadas por incêndio ou queimada perto da rede elétrica.

É importante observar a sobreposição natural existente entre cada tipo de causas apontadas acima (1 à 9). Causas como *Vegetal* ou *Interferência do Meio Ambiente* podem facilmente ser confundidas em uma inspeção visual. O mesmo acontece com *Falha no Componente* (Chaves fusíveis) e *Sobrecarga*. Assim, salienta-se ainda mais a dificuldade de coletar informações sobre desligamentos involuntários. Portanto, uma base de dados ao ser desenhada para o objetivo de suportar sistemas de apoio a decisão ou de determinação automática de prováveis causas de desligamentos, deve possuir, no mínimo, o conjunto de informações descritas anteriormente, com mecanismos que possam evitar as sobreposições de causas.

Como exemplo da organização da coleta e armazenagem de dados de eventos de desligamentos não programados, cita-se as referências Pretto *et alii* (2003), Pretto *et alii* (2005a), Pretto (2005b) e Pretto *et alii* (2006). Nessas referências os eletricistas utilizam um software desenvolvido para computadores móveis na coleta de informações a respeito do local do evento de desligamento. Ao final do dia de trabalho e do atendimento de algumas ocorrências, os dados coletados são enviados ao computador do centro de operação e armazenados em um banco de dados. Essa base de dados pode servir como fonte de informação para estudos estatísticos e análises probabilísticas, permitindo a correta identificação das causas mais prováveis de um desligamento não programado.

Entretanto, se a empresa não tem um sistema projetado e direcionado para um sistema de identificação automática de causa de eventos não programados, as bases de dados irão apresentar variáveis inconsistentes e com elevado grau de incerteza para tratar adequadamente este tipo de abordagem. Isso é causado principalmente pela ausência de uma metodologia adequada na coleta dos dados no local do evento de desligamento e, também, pelas próprias características dos sistemas de energia, que apresentam muitas variáveis difíceis de serem observadas.

A base de dados utilizada, contendo informações sobre o histórico de eventos de desligamentos, é a disponível na empresa, onde os dados brutos coletados pelos eletricistas não sofreram nenhum tipo de tratamento. Esta situação permite concluir que, geralmente, as informações não estão organizadas para a utilização direta de sistemas computacionais. A próxima seção apresenta a metodologia proposta para contornar essa situação e criar um conjunto de dados adequados.

4 METODOLOGIA DESENVOLVIDA

A metodologia desenvolvida neste artigo tem por objetivo tratar informações a partir de bases de dados pré-existentes, isto é, bases de dados que não foram projetadas com um fim específico e orientadas para a identificação de causas de desligamentos forçados.

Para o desenvolvimento, foi utilizada uma base de dados real, na qual são armazenados milhares de ocorrências vinculadas a desligamentos não programados na rede de distribuição de energia. Após analisar cuidadosamente os campos e o conteúdo da base disponível, verificou-se que a mesma apresentava campos incompletos, informações contraditórias, campos sem dados e muitos outros problemas. Esses fatores colaboraram para concluir que a base, no estado inicial, era inapropriada para a utilização direta.

As próximas seções descrevem as etapas que envolvem o processo de *KDD* utilizado na metodologia. A descrição do trabalho segue a lógica do desenvolvimento aplicada a este artigo, onde algumas decisões de cunho prático que envolve experiência de campo são consolidadas junto a profissionais que atuam diretamente na rede ou na operação de sistemas de distribuição. Algumas etapas são mais pormenorizadas devido a sua importância no processo geral. Um exemplo pode ser observado na etapa de mineração de dados, a qual será apresentada na seção 4.5 com uma descrição mais detalhada.

4.1 Definição da Nova Estrutura de Dados

A etapa de definição das variáveis, correspondente a um pré-processo de *KDD*, está amparada na metodologia desenvolvida por Pretto *et alii* (2006). Neste artigo as interrupções com origem no sistema de distribuição são identificadas mediante uma classe que representa o possível evento ocorrido e que gerou a saída forçada de algum equipamento e/ou a interrupção do fornecimento de energia. Nessa abordagem, um evento ocorrido na rede é analisado através das condições de operação que envolve a interrupção, identificando-se uma causa através do apontamento dos seguintes fatores:

- Elementos de Interrupção

- Clima
- Vento
- Fatos Associados
- Vegetação
- Objetos

Geralmente, em empresas de distribuição, as interrupções são armazenadas com apontamentos diretos, seguindo um padrão de identificação de causa pré-definido. Dessa forma, o trabalho de definição das variáveis contou com a tentativa de associar a cada interrupção as variáveis apontadas acima. Embora a literatura técnica apresente algumas classificações mais diretas, como em Oprisan *et alii* (1991), onde o autor desenvolve um sistema para armazenamento de dados sobre interrupção de energia para cálculo do desempenho da rede, optou-se por utilizar a classificação de Pretto (2005) .

Sendo assim, cada observação bruta da base de dados foi reformatada para o modelo acima descrito. Porém, esta tarefa não é possível de ser realizada diretamente, e algumas inferências são necessárias a fim de ajustar os dados. As regras de manipulação dos dados e as regras de inferência são descritas a seguir.

Resumidamente, o objetivo da metodologia proposta consiste em gerar regras que formatem os dados brutos na classificação descrita anteriormente.

4.2 Seleção dos Dados

No primeiro passo da utilização da técnica de *KDD*, conforme indicado na Figura 1, os autores realizaram a seleção manual dos campos de maior interesse oriundos do banco de dados original (ver Tabela 1), ordenados através de um critério que reflete a importância de cada campo no auxílio ao processo de mineração dos dados. Os campos selecionados para o estudo estão representados de forma simplificada na Tabela 1:

Tabela 1: Campos selecionados do banco de dados original

NUM_1	X_CORD	Y_CORD	DEV_TYPE_NAME	CLIMA	CAUSA

Onde cada campo pode ser descrito como segue:

NUM_1: Número da Ordem de Serviço do desligamento;

X_CORD: Coordenada X (*Universal Transverse Mercator UTM*) do local onde ocorreu a interrupção;

Y_CORD: Coordenada Y (*Universal Transverse Mercator UTM*) do local onde ocorreu a interrupção;

DEV_TYPE_NAME: Equipamento (s) que foi(ram) afetado (s) na interrupção;

CLIMA: Clima no momento da interrupção;

CAUSA: Causa apontada segundo a coleta do eletricista em campo, seguindo o padrão da empresa.

4.3 Limpeza e Integração dos Dados

De acordo com a Figura 1, a etapa de limpeza e integração dos dados forma a primeira fase do processo de *KDD*. Utilizando o histórico de eventos fornecido por uma distribuidora de energia situada na região sul do Brasil, identificou-se a estrutura relacional aplicada ao banco de dados através do seu diagrama de entidades e relacionamentos (ER), e verificaram-se os tipos de ferramentas necessárias para proceder à etapa em questão. As consultas e manipulações do banco de dados foram realizadas utilizando a linguagem *Structured Query Language* (*SQL*). Dessa forma, foi possível extrair da estrutura original as variáveis definidas na seção anterior, abandonando a estrutura original e adotando uma nova estrutura relacional, adequada ao armazenamento das variáveis de interesse. Com o objetivo de pormenorizar esse trabalho, apresenta-se na Figura 2 um exemplo descritivo de uma seqüência de filtros criados com o objetivo de limpar e integrar as variáveis de interesse na nova estrutura utilizada.

Assim, do conjunto original de eventos de desligamentos armazenados no banco de dados, constatou-se que apenas uma pequena porcentagem desses dados era adequado. Portanto, grande parte desses registros foi descartado da nova base de dados. Conforme a Figura 2, os primeiros registros a serem eliminados da base de eventos foram os que apontavam como causa um desligamento que havia sido programado pela companhia distribuidora de energia, cita-se, desligamentos para manutenção preventiva, alteração para melhorias, desligamento a pedido do cliente, entre outros. Em seguida, foram retirados os registros contendo os valores inexistentes na lista de códigos de causa da companhia e valores do tipo *NULL*, indicando que o campo não foi preenchido pelo eletricista na hora do atendimento da ordem de serviço.

Dessa forma, foi possível separar apenas registros de desligamentos não-programados, os quais interessam ao estudo em questão. Porém, havia ainda a necessidade de limpar e integrar este novo conjunto de dados, o qual apresentava problemas, tais como: situações contraditórias - como exemplo: a indicação da causa *Descarga Atmosférica* em registro de *clima bom*. Especialmente nesses campos adotou-se o registro *clima bom* como parâmetro *verdade*, principalmente devido ao cruzamento de outras interrupções nas mesmas datas de ocorrência que também registravam o *clima bom*. O registro do *clima* causou outras preocupações durante o pro-

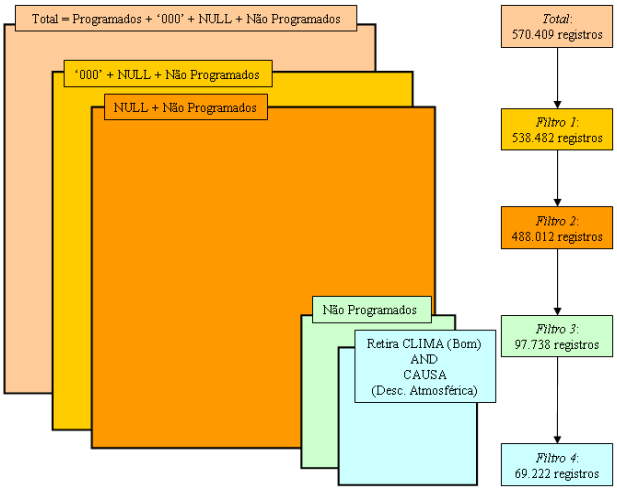

Figura 2: Limpeza e Integração de dados

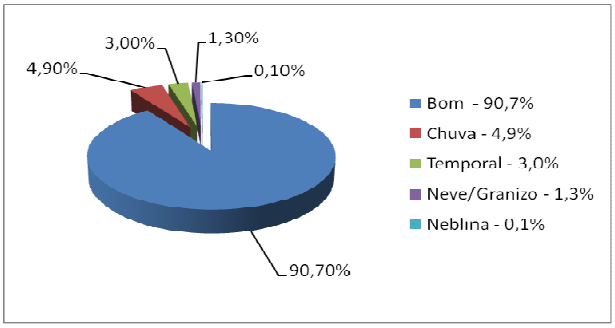

Figura 3: Desligamentos considerados válidos

cesso de limpeza e integração dos dados , que serão expostas ainda nesta seção, mas que não afetaram as decisões até o momento.

Nesse ponto, buscou-se identificar e remover registros com contradições. A Figura 2 apresenta uma visão da ordem de grandeza (quantificação) de dados eliminados do conjunto inicial utilizando os critérios descritos acima. Após essa limpeza e integração, dos 570.409 registros fornecidos, apenas 12% dos dados foram considerados aderentes aos eventos registrados para serem utilizados no sistema, totalizando 69.222 registros. O número reduzido de dados, já filtrados, presentes nesta base de dados (em torno de 12%) ressalta a importância de um processo qualificado para a coleta de informações desses eventos. Em Pretto *et alii* (2006) são expostos alguns exemplos de sistemas de coleta de dados de interrupções, em que a qualidade da informação é o foco principal. Neste ponto, é importante elaborar uma reflexão acerca dos registros coletados em campo. Uma das maiores preocupações de Pretto *et alii* (2006) foi retirar a responsabilidade de apontamento de uma causa de interrupção do eletricista (coletor), pois observou-se que durante a recomposição da rede, nem sempre é possível identificar diretamente uma causa e, na grande maioria das vezes, os eletricistas são direcionados para recompor rapidamente a interrupção, com o objetivo de mitigar indicadores de duração e frequência, sendo secundária a tarefa de apontamento da causa.

Promovendo uma nova análise dos dados após a primeira fase de Limpeza e Integração, observou-se que grande parte das interrupções foram registradas com o campo "CLIMA" preenchido como "bom". A Figura 3 apresenta essa característica. Após examinar cuidadosamente os registros, e verificar os procedimentos de operação seguidos pela empresa, constatou-se que a nova divisão continha uma tendência registrada, ou seja, durante a recomposição das redes, o campo "CLIMA" da ficha de recomposição preenchida pelo eletricista era um campo ignorado, sendo posteriormente preenchido com seu valor *padrão*. Obviamente, essa análise deveria considerar essa condição.

O fato verificado na nova base de dados contraria uma afirmação de Brown (2002), onde o autor afirma que em clima adverso as interrupções são mais freqüentes que em clima bom, para a grande maioria das empresas de distribuição de energia. Com base nessa referência e nos procedimentos de operação da distribuidora, constatou-se que a maioria dos registros de interrupções de interesse a este trabalho, contidos na base de dados, deveria ocorrer em clima adverso, fato que conduziu a uma correção da tendência verificada e apresentada na Figura 3. Assim, com intuito de tornar a base de dados mais fiel à realidade da operação dos sistemas de distribuição, foi realizada mais uma fase de Limpeza e Integração dos Dados, com a adição dos seguintes critérios:

- 70% da nova base são compostas por informações pertencentes às condições de clima adverso, conforme apresentado na Figura 4;
- 30% da nova base são compostas por eventos com clima bom.

Os dados filtrados foram divididos em um conjunto contendo eventos ocorridos durante clima bom, com registros não tendenciosos e outro conjunto em clima adverso, como ilustra a Figura 4.

Dessa forma, a redução proposta para a base de dados traduz uma forte aderência aos aspectos de identificação das variáveis definidas para o problema. Os registros escolhidos para compor o conjunto "Clima Bom" não apresentam nenhum indicativo que se oponha à classificação, como por exemplo, a presença de vento ou neblina nos registros.

A nova base foi acomodada em uma estrutura com 8.888

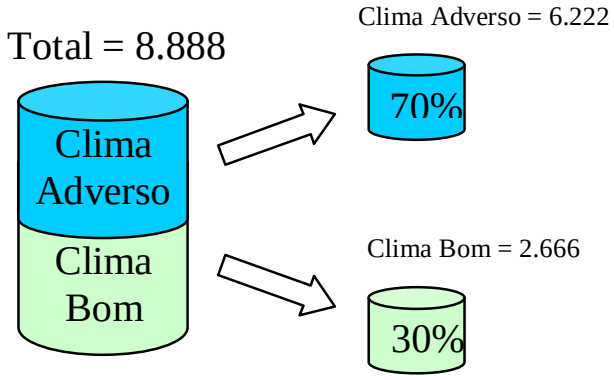

Figura 4: Conjuntos de desligamento em clima bom e clima adverso

eventos registrados. É importante observar que a base iniciou com 570.409 registros e passou para 8.888 eventos de interrupções com condições de utilização, isto é, somente 1,5% foram aproveitados.

4.4 Transformação dos Dados

A Tabela 2 apresenta os dados em seu estado bruto sem nenhum processo de qualificação.

Tabela 2: Formato de dados brutos

DEV_TYPE_ NAME	CLIMA	CAUSA	X_CORD	Y_CORD
Chave Fusível	Temporal	Poluição	40227267	679542682
Transformador	Chuva	Poluição	57937709	677247941
Chave Fusível	Temporal	Descarga Atmos-férica	46971602	676555018
Chave Fusível	Temporal	Descarga Atmos-férica	56973869	678813192

Após a transformação dos dados da Tabela 2, utilizando-se a classificação proposta na seção 3.1, será originada uma estrutura de dados como descrita na Tabela 3.

Tabela 3: Estrutura de dados

ELEMENTOS	FATOS	CLIMA	VENTO	OBJETOS	VEGETAÇÃO	CAUSAS

A seção 4.5.3 descreve em maiores detalhes esse processo.

4.5 Mineração de Dados

A etapa de Mineração de Dados consiste em extração de informação interessante, não trivial, implícita, previamente desconhecida e potencialmente útil nas informações armazenadas em grandes massas de dados (Fayyad *et alii*, 1996). Pode-se utilizar diferentes formas para produzir padrões extraídos através da mineração de dados, e cada uma irá ditar o tipo de técnica que será utilizada para produzir a estrutura de saída dos dados. Para essa tarefa existe uma grande variedade de técnicas que costumam ser utilizadas, tais como regras de associação, regras de produção, padrões de seqüências, agrupamento (*clustering*) e padrões em séries temporais. Neste trabalho, a extração de padrões (conhecimento) foi realizada através de regras de classificação (Norvig e Russel, 2003), tendo em vista o formato de representação de conhecimento desejado para a formação do banco de dados necessário ao treinamento de uma Rede Bayesiana. Para uma melhor compreensão desta etapa, apresenta-se a seguir, em maior detalhe, as decisões que formam as regras de classificação utilizadas neste trabalho.

4.5.1 Extração de Padrões

Os primeiros sistemas baseados em conhecimento foram sistemas utilizando regras (Buchanan e Shortliffe, 1984). Nesses sistemas, o processo de tomada de decisão humano é modelado por regras do tipo "*se* P *então* Q", simbolicamente $P \rightarrow Q$. Portanto, as regras podem expressar relacionamentos lógicos e equivalências de definições para simular o raciocínio humano (Araribóia, 1987).

A escolha da tarefa a ser executada no processo é feita de acordo com os objetivos desejáveis para a solução a ser encontrada. Os dois principais objetivos da mineração de dados são:

- Predição: envolve o uso de algumas variáveis nas bases de dados para predizer valores futuros ou desconhecidos de outras variáveis de interesse.
- Descrição: busca obter padrões que descrevam os dados e aprender uma hipótese generalizada, um modelo, a partir dos dados selecionados.

A Figura 5 mostra os tipos de tarefa que podem ser executadas.

Atividades de predição consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo desconhecido. Por exemplo, sabendo que a causa de um desligamento foi *Vegetal* pode-se inferir com um nível de incerteza que a área afetada possui vegetação abundante e,

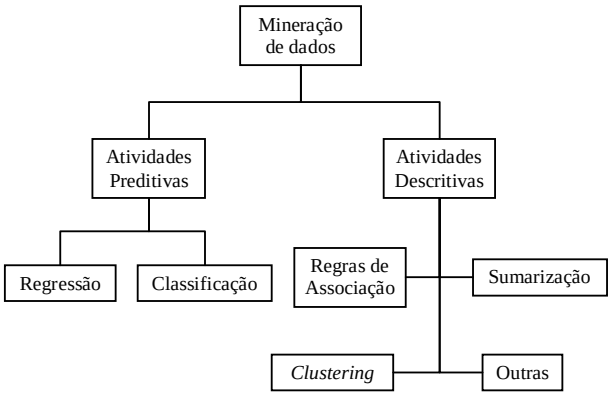

Figura 5: Tarefas de Mineração de Dados

portanto, pode-se atribuir o estado *Não Podado* para o campo *vegetação*, mesmo que esse campo não esteja presente no banco de dados original. Ou ainda, sabendo que a causa original é *Acidente*, isso dará origem a um estado *Acidente* no novo campo *Fatos Associados*.

Neste trabalho, criou-se um conjunto de regras baseadas no conhecimento de especialistas sobre o domínio, tendo como objetivo predizer variáveis que estão implícitas no banco de dados original, mas que podem ser extraídas ou inferidas utilizando técnicas de mineração de dados. Através desse procedimento criou-se uma base de conhecimento mais fiel ao domínio em estudo.

4.5.2 Classificação Baseada em Regras

Dada a característica do problema a ser resolvido, optou-se pela adoção da técnica de classificação baseada em regras, uma vez que o conhecimento do especialista pode ser incorporado para a formação e verificação destas regras. Esta é uma técnica para classificar dados que utiliza um conjunto de regras do tipo "*se... então...*". As regras, para o modelo, são representadas de forma disjunta, ou seja, somente uma regra é disparada por vez, $R = (r_1 \vee r_2 \vee ...r_k)$, onde R é o conjunto de regras e r_i são as regras de classificação (Han e Kamber, 2000).

4.5.3 Motor de Inferência da Mineração de Dados

Com o objetivo de obter-se uma boa base de conhecimento a partir de uma fonte de informação pobre, os dados foram tratados utilizando um conjunto de regras de classificação, a qual se denomina motor de inferência do sistema. A Figura 6 mostra a informação na entrada da máquina de inferência e a informação desejada na saída, utilizada neste trabalho, onde:

1. ELEMENTOS: elemento em que ocorreu a falha;

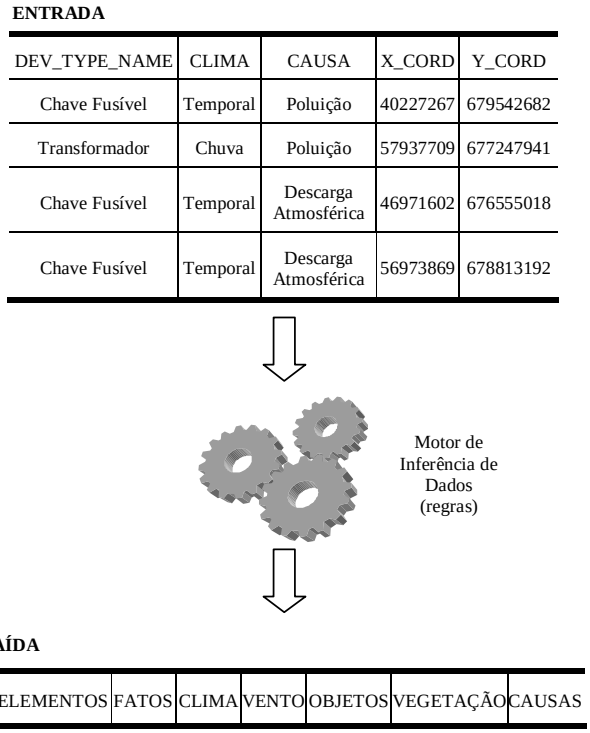

Figura 6: Formato dos dados na entrada e na saída do motor de inferência da mineração de dados.

2. FATOS: fatos associados ao evento de desligamento que foram observados no local;
3. CLIMA: condições climáticas no momento do desligamento;
4. VENTO: intensidade do vento no momento do desligamento;
5. OBJETOS: se existem objetos presos aos condutores;
6. VEGETAÇÃO: condições da vegetação;
7. CAUSA: causa provável apontada no desligamento.

O nome dos campos foi reduzido com o propósito de diminuir o espaço utilizado para formalização das regras, como mostram as Tabelas 4 e 5.

Tabela 4: Legenda para os campos de entrada.

DEV_TYPE_ NAME	CLIMA	CAUSA	X_CORD	Y_CORD
DEV	CLe	CSe	-	-

Tabela 5 – Legenda para os campos de saída.

Tabela 5: Legenda para os campos de saída.

ELEMENTO	FATOS	CLIMA	VENTO	OBJETOS	VEGETAÇÃO	CAUSAS
EL	FT	CLs	VT	OBJ	VEG	CSs

A seguir será descrito o conjunto de regras utilizado para classificação dos dados.

4.5.3.1 Campo ELEMENTOS

Utilizando o conceito de Elemento de Interrupção, baseado na classificação apresentada em Pretto (2005b), que indica que tipo de equipamento ou estrutura está envolvida na interrupção, e o campo 'CAUSA' presente no histórico de eventos original, é possível concluir que elemento físico da rede foi afetado para causar o desligamento. Assim, baseado em informações sobre a causa do evento e nas classes de elemento afetadas durante uma falha no fornecimento de energia (Brown, 2002), um novo campo chamado *elementos* foi criado. Portanto, para cada evento de desligamento analisado no banco de dados original, uma nova classe de *elementos* foi atribuída de acordo com os dados disponíveis nesse banco de dados. Todos esses *elementos*, que indicam que estrutura ou equipamento estava diretamente relacionado com o desligamento forçado, foram classificados com o auxílio de um especialista. Para cada conjunto de informações oriundas da base de dados original, um elemento de interrupção foi escolhido, como demonstra o exemplo abaixo.

0. ***Elemento de Interrupção***: ***Informação BD Original;***

Classificação:

1. **Poste**: Poste Podre, Poste, Abalroamento de Poste, Erosão;
2. **Equipamentos**: Religador, Descarga Atmosférica, Animais/Pássaros, Transformador de Potência (TP), Ferragens, Pára-Raios, Vandalismo, Regulador, Transformador, Elo Fusível Queimado, Chave Fusível, Chave Faca;
3. **Condutores**: Conexões (todo ponto de ligação), Ramal de Ligação, Condutor Desregulado, Condutor, Vento, Emenda, Amarrilho, Animais/Pássaros;
4. **Isoladores**: Isoladores, Animais/Pássaros;
5. **Cruzeta**: Cruzeta.

Os elementos que não foram classificados diretamente a partir da sua causa original passaram por um processo de classificação através de regras, que constitui um agrupamento dos campos DEV_TYP_NAME e CSe apresentados na tabela 4, originando informação para o novo campo Elementos.

Definiram-se, então, um conjunto de 22 regras para relacionar essas causas a um elemento de interrupção. Com o objetivo de ilustrar essa etapa, apresentam-se a seguir algumas dessas regras:

r_1 : (CSe = vegetal) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos
r_2 : (CSe = vegetal) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores
r_3 : (CSe = sobrecarga) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos
r_4 : (CSe = sobrecarga) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores
r_5 : (CSe = queimada / incêndio) $\wedge$ ...
[(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

Através do conjunto de 22 regras criadas para os casos em que o elemento de interrupção não pode ser extraído diretamente, é possível identificar qual equipamento ou estrutura mais provavelmente estava envolvido no desligamento.

4.5.3.2 Campo FATOS

Este campo contém informações sobre atividades que são realizadas próximas à rede elétrica e que podem influenciar em um desligamento.

Com base nas causas apontadas no conjunto de dados utilizado como entrada para as regras foi possível tirar algumas conclusões a respeito de fatos associados ao desligamento. Abaixo são apresentados exemplos de algumas regras que foram adotadas:

r_1 : (CSe = queimada / incêndio) $\rightarrow$ queimada / incêndio
r_2 : (CSe = danos causados por terceiros) $\rightarrow$ vandalismo
r_3 : (CSe = abalroamento de poste) $\rightarrow$ acidente
r_4 : (CSe = inundações) $\rightarrow$ inundação
r_5 : (CSe = erosão) $\rightarrow$ erosão

As regras associam os dados contidos no seu estado bruto à nova classificação de FATOS. Pode-se considerar como saída do motor de inferência para este campo os seguintes estados:

Estados acidente, queimada/incêndio, empresa, vandalismo, inundação e erosão.

Para qualquer outra causa apontada atribuiu-se ao campo

"FATOS" o símbolo '*', indicando que a variável não foi observada naquele evento.

4.5.3.3 Campo CLIMA

No banco de dados original, *Vento* era um dos estados que a variável *CLIMA* podia assumir. No entanto, caso estivesse chovendo e ventando ao mesmo tempo, não seria possível descrever adequadamente as condições climáticas do local do evento. Por isso, o estado *Vento* foi eliminado da variável *CLIMA*, e uma nova variável *VENTO* foi criada.

Assim, criou-se a seguinte regra, com o objetivo de substituir o antigo estado *Vento* por um novo estado no campo *CLIMA*:

$$r_1: \ (\text{CLe} = \text{vento}) \rightarrow \text{bom}$$

Os demais estados dessa variável foram extraídos diretamente a partir do banco original, pois já estavam em um formato adequado para o estudo realizado.

Estados bom, chuva, temporal, neve/granizo, neblina.

4.5.4 4.5.3.4 Campo VENTO

A definição do estado desse campo baseia-se no que pode ser observado no campo *CLIMA*, já que as condições climáticas dão uma idéia sobre a intensidade do vento. Por exemplo, num dia de tempestade, a chance de se ter ventos fortes é alta, então se pode dizer que num dia de tempestade, o *vento* é excessivo. A partir disso derivaram-se as seguintes regras para definir o estado da variável.

$$\begin{aligned} r_1: &\ (\text{CLe} = \text{bom}) \rightarrow \text{sem vento} \\ r_2: &\ (\text{CLe} = \text{chuva}) \rightarrow \text{moderado} \\ r_3: &\ (\text{CLe} = \text{tempestade}) \vee (\text{CLe} = \text{vento}) \vee \ldots \\ &\ (\text{CLe} = \text{granizo}) \rightarrow \text{excessivo} \end{aligned}$$

Estados sem vento, moderado, excessivo.

4.5.5 4.5.3.5 Campo OBJETOS

É comum a ocorrência de desligamentos ocasionados por objetos presos na rede elétrica, portanto, essa variável deve ser observada no local do desligamento. Nesse caso é indicada somente a presença ou não de objetos presos na rede.

O estado desse campo foi definido a partir da causa apontada nos dados de entrada. A seguinte regra descreve o processo:

$$\begin{aligned} r_1: &\ (\text{CSe} = \text{vegetal}) \vee (\text{CSe} = \text{pandorga}) \ \vee \ldots \\ &\ (\text{CSe} = \text{animais / pássaros}) \ \vee \ldots \\ &\ (\text{CSe} = \text{objetos estranhos na rede}) \ \vee \ldots \\ &\ (\text{CSe} = \text{bola na rede elétrica}) \rightarrow \text{sim} \\ r_2: &\ (\text{CSe} \neq \text{vegetal}) \wedge (\text{CSe} \neq \text{pandorga}) \ \wedge \ldots \\ &\ (\text{CSe} \neq \text{animais / pássaros}) \ \wedge \ldots \\ &\ (\text{CSe} \neq \text{objetos estranhos na rede}) \ \wedge \ldots \\ &\ (\text{CSe} \neq \text{bola na rede elétrica}) \rightarrow \text{não} \end{aligned}$$

Aplicando as regras, a variável irá assumir um dos dois estados definidos para este campo:

Estados sim, não.

4.5.3.6 Campo VEGETAÇÃO

De acordo com Brown (2002), contato com árvores é uma das três causas mais comuns de interrupção de energia. À primeira vista, nenhuma conclusão a respeito das condições da vegetação pode ser tirada dos dados brutos. No entanto, se as informações disponíveis sobre a causa e a coordenada geográfica do local do desligamento forem cruzadas, pode-se inferir algo sobre o estado da vegetação para o evento, como, por exemplo, estado *podado* ou *não-podado*.

Os eventos de desligamento não-programados foram agrupados por municípios utilizando o algoritmo *k-means* (MacQueen, 1967) e (Teknomo, 2006). A idéia deste algoritmo é classificar em agrupamentos (*clusters*) as informações com características similares, baseada na análise comparativa entre as informações e valores numéricos dos dados, de forma automática e sem a necessidade de uma pré-classificação. Por causa desta característica, o *k-means* é considerado como um algoritmo de mineração de dados não supervisionado.

Estados podada, não podada.

O funcionamento do algoritmo *k-means* é bastante simples. O número de classes ou agrupamentos deve ser informado no algoritmo. Os passos do algoritmo são descritos brevemente a seguir, apoiados pela Figura 7.

Os parâmetros de entrada da função do algoritmo são os seguintes:

- m – matriz de dados que contém os pontos correspondentes às coordenadas geográficas X e Y (*UTM*) dos eventos de desligamento;
- k – número de grupos (*clusters*), nesse caso é 272, que representa o número de municípios de atuação da empresa. É importante salientar que os centróides de entrada do algoritmo correspondem as coordenadas (*UTM)* centrais dos municípios;

Desligamentos causados por vegetal

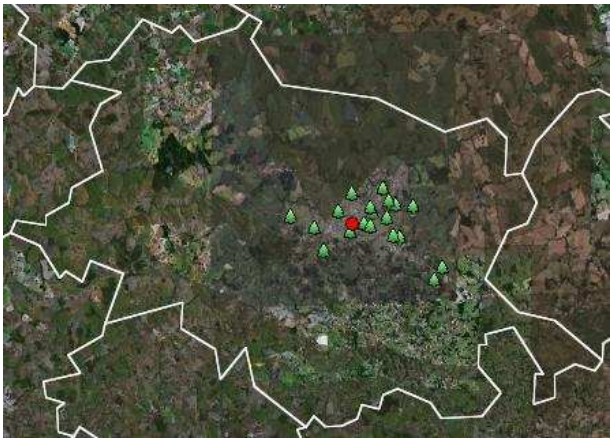

Figura 7: Ilustração para o campo vegetal

- israd – matriz de dados que contém os pontos correspondentes às coordenadas geográficas X e Y (*UTM*) dos 272 municípios, que representam os centróides.

A função retorna uma matriz de dados idêntica à matriz m, porém com uma coluna adicional que representa os grupos (municípios) onde cada evento de desligamento ocorreu. Dessa forma, foi possível saber quantos desligamentos causados por vegetal ocorreu em cada município, permitindo a extração de dados estatísticos que serão utilizados como critério na definição do estado da vegetação no local do evento, como demonstra a Tabela 6.

Tabela 6: Dados estatísticos extraídos

Desligamento Causa: Vegetal	N.o de Municípios	Média μ	Desvio Padrão σ	Variância σ^2
1759	272	6,47	4,98	24,76

O processo de mineração de dados passou então a ter mais uma fonte de informação, o campo *número de desligamentos por município*. Esse número de desligamentos por município foi utilizado na criação da regra que define se aquela região onde ocorreu o evento possui uma vegetação abundante ou não. A Tabela 7 mostra como ficou a nova base contendo os novos campos correspondentes.

A regra leva em consideração o número médio de desligamentos causados por vegetal em cada município da concessão da distribuidora. A média μ = 6,47 foi truncada para μ = 6. Assim, sabendo que um evento ocorreu num determinado município e que nesse município ocorreu um número x de desligamentos causados por vegetal, é possível concluir

Tabela 7: Nova base de dados após aplicação do *k-means*

Evento	DEV_TYPE_ NAME	CLIMA	CAUSA	ÁREA	Nº VEGETAIS
1	Chave Fusível	Chuva	Descarga Atmosférica	1	9
2	Transformador	Chuva	Descarga Atmosférica	1	9
3	Transformador	Chuva	Descarga Atmosférica	1	9
...	...	...	...	...	...
2420	Transformador	Vento	Ramal de Ligação	67	4
2421	Transformador	Bom	Conexões	67	4
2422	Chave Fusível	Bom	Vegetal	67	4
...	...	...	...	...	...
6204	Chave Fusível	Temporal	Poste Podre	173	21
6205	Chave	Temporal	Poste Podre	173	21
6206	Transformador	Temporal	Poste Podre	173	21
6207	Transformador	Temporal	Transformador	173	21
6208	Chave Fusível	Vento	Descarga Atmosférica	173	21
6209	Transformador	Vento	Descarga Atmosférica	173	21
...	...	...	...	...	...
8885	Chave Fusível	Bom	Animais Pássaros	272	0
8886	Transformador	Bom	Vento	272	0
8887	Chave Fusível	Bom	Chave Fusível	272	0
8888	Transformador	Bom	Animais Pássaros	272	0

algo a respeito do estado da vegetação. Neste caso a regra elaborada foi descrita assim:

$$r_1 : (x > \mu) \rightarrow \text{não podado}$$
$$r_2 : (x < \mu) \rightarrow \text{podado}$$

4.5.3.7 Campo CAUSA

Este é o campo que deve ser preenchido com a causa do desligamento. Portanto, este é o campo que será preenchido pela aplicação da metodologia proposta. As causas presentes no banco de dados original foram classificadas dentro de uma das nove classes de causas de desligamento utilizadas nesse trabalho:

- falha no componente

- sobrecarga
- clima adverso
- descarga atmosférica
- interferência do meio-ambiente
- vegetal
- interferência humana
- acidente
- queimada/incêndio.

Com o objetivo de preencher o campo “CAUSAS” da nova base de dados, utilizou-se um novo conjunto de regras, as quais são exemplificadas a seguir.

r_1 : (CSe = vegetal) $\rightarrow$ vegetal
r_2 : (CSe = descarga atmosférica) $\rightarrow$ descarga atmosférica
r_3 : (CSe = condutor desregulado) $\vee$ (CSe = condutores) $\vee$ (CSe = poste podre) $\vee$ ...
(CSe = poste) $\vee$ (CSe = ferragens) $\vee$ (CSe = regulador) $\vee$ ...
(CSe = transformador) $\vee$ (CSe = TP - transformador de potência) $\vee$ ...
(CSe = TC - transformador de corrente) $\vee$ ...
(CSe = TP - transformador de potência) $\rightarrow$ falha do componente
r_4 : (CSe = vento) $\rightarrow$ clima adverso
r_5 : (CSe = sobrecarga) $\rightarrow$ sobrecarga
r_6 : (CSe = animais/pássaros) $\vee$ (CSe = poluição) $\rightarrow$ interferência do meio
r_7 : (CSe = queimada/incêndio) $\rightarrow$ queimada/incêndio
r_8 : (CSe = abalroamento de poste) $\rightarrow$ acidente
r_9 : (CSe = danos causados por terceiros) $\vee$ (CSe = vandalismo) $\vee$ ...
(CSe = pandorga) $\vee$ (CSe = objetos estranhos na rede) $\vee$ ...
(CSe = redes de comunicação) $\vee$ (CSe = bola na rede elétrica) $\vee$ ...
(CSe = falta de manutenção) $\rightarrow$ interferência humana
r_{10} : (CSe = conexões) $\vee$ (CSe = isolador) $\vee$ (CSe = ramal de ligação) $\vee$ ...
(CSe = pára - raios) $\vee$ (CSe = religador) $\vee$ (CSe = cruzeta) $\vee$ ...
(CSe = emenda) $\vee$ (CSe = elo fusível queimado) $\vee$ (CSe = amarrilho) $\vee$ ...
(CSe = chave fusível) $\vee$ (CSe = chave faca) $\wedge$ [(CLe = temporal) $\vee$ ...
(CLe = vento) $\vee$ (CLe = granizo)] $\rightarrow$ clima adverso

5 RESULTADOS

O principal objetivo desse trabalho foi aplicar a técnica de *KDD* para a extração de conhecimento de uma base de dados bruta relacionada à interrupção de energia em redes de distribuição. Uma fase importante foi a relacionada à etapa de mineração de dados, onde o processo consistiu em extração de informações úteis relacionadas aos aspectos que envolvem uma interrupção não programada. Entre as informações interessantes deste tipo de desligamento, destacam-se as informações não triviais vinculadas à associação de condições climáticas da hora das interrupções, informações implícitas vinculadas aos apontamentos incompletos dos coletores (eletricistas) e as informações previamente desconhecidas vinculadas aos aspectos operacionais dos sistemas de distribuição. Um exemplo do resultado do processo de *KDD* pode ser visto na Tabela 8, onde uma nova base de dados foi composta para ser utilizada como entrada para um aplicativo que identifica causas de desligamentos mediante algumas observações de campo.

Tabela 8: Resultado do processo de *KDD* conhecimento extraído para treinamento de sistemas especialistas

ELEMENTO	FATOS	CLIMA	VENTO	OBJETO	VEGETAÇÃO	CAUSAS
Equipamentos	*	Chuva	Moderado	Não	SemPoda	DescAtm
Poste	Acidente	Bom	SemVento	Não	SemPoda	Acidente
Condutores	*	Bom	SemVento	Sim	SemPoda	Vegetal
...	...	...	...	...	...	...
Equipamentos	*	Temporal	Excessivo	Não	Podada	DescAtm
Isoladores	*	Chuva	Moderado	Não	SemPoda	FalhaComp
Equipamentos	*	Chuva	Moderado	Não	SemPoda	DescAtm
...	...	...	...	...	...	...
Condutores	*	Bom	SemVento	Sim	SemPoda	Vegetal
Poste	Acidente	Bom	SemVento	Não	Podada	Acidente
Equipamentos	Empresa	Bom	SemVento	Não	Podada	InterfHum
...	...	...	...	...	...	...
Condutores	*	Chuva	Moderado	Não	SemPoda	FalhaComp
Equipamentos	Vandalismo	Bom	SemVento	Não	SemPoda	InterfHum
Isoladores	*	Chuva	Moderado	Não	SemPoda	FalhaComp

Como ilustração desse resultado, a Figura 8 apresenta uma Rede Bayesiana (Tronchoni *et alii*, 2007) treinada a partir da base de conhecimento apresentada na Tabela 8.

De acordo com algumas variáveis observadas no local da interrupção, as prováveis causas do corte de fornecimento de energia são identificadas pela rede Bayeseana, apoiando os operadores na tomada de decisão para alocação de equipes e recursos na área de concessão da empresa.

5.1 Resultados Quantitativos

O método da prova bipartida (split-half method) (Congdon, 2003) (foi aplicado na validação da rede. Ele consiste em dividir o conjunto de dados em duas partes iguais, uma para treinamento e uma para validação. O conjunto obtido após o tratamento dos dados é chamado de conjunto original e totaliza 8888 amostras. A partir do conjunto original, foram selecionados aleatoriamente seis conjuntos contendo 4444 amostras cada, formando três pares de conjuntos de treinamento e validação.Para cada evento existem variáveis ou nós que descrevem as condições no local de desligamento, e um nó que indica as possíveis causas de um desligamento forçado. Os estados observados em cada evento do conjunto de validação

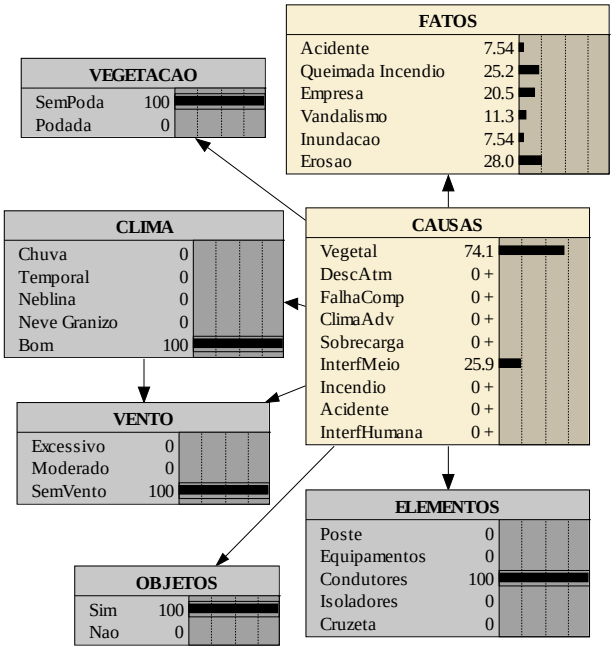

Figura 8: RB treinada e instanciada com os estados observados nas variáveis de um evento

são instanciados em seus respectivos nós, com exceção do nó 'CAUSAS', que irá indicar a causa provável do desligamento. Para cada evento, a rede apontou uma causa. Essa causa estimada é comparada com a causa real.

O processo se repetiu para todos os eventos da amostra. Assim foi possível calcular o número de vezes que a rede teve sucesso no diagnóstico da causa e o número de vezes em que falhou. Foi possível determinar o erro da Rede Bayesiana dividindo-se o número de vezes em que o diagnóstico falhou pelo número total de eventos, como mostra a equação 1

$$erro = \frac{\text{num. de diagnosticos errados}}{\text{num. total de eventos}} \qquad (1)$$

A Tabela 9 apresenta os resultados obtidos através das 3 redes Bayesianas treinadas utilizando os 3 conjuntos de dados de entrada (chamadas de RB1, RB2 e RB3). É importante salientar que cada rede apresenta os resultados do próprio conjunto de treinamento (erro mais baixo) e mais 2 resultados dos conjuntos restantes.

5.2 Desenvolvimento Computacional

As etapas do método *KDD* foram implementadas utilizando a liguagem *SQL* em um banco de dados *Oracle*. As regras de inferência e o algoritmo *k-means* foram implementados utilizando a ferramenta de programação *C++ Builder 6* da

Tabela 9: Erro dos treinamentos da rede.

	Conjunto 1	Conjunto 2	Conjunto 3	M	σ
RB1	9.60%	8.50%	9.00%	9.00%	0.56
RB2	9.20%	8.20%	8.70%	9.70%	0.50
RB3	8.80%	8.20%	8.50%	8.50%	0.30
Erro Médio	9.20%	8.30%	8.73%	8.74%	0.45

Borland. A rede Bayeseana foi implementada com o auxílio do aplicativo *Bayes Net Toolbox for Matlab (Murphy, K.P., 2001)*.

Os tempos de execução dos filtros e dos algoritmos podem ser considerados irrelevantes, apesar do tamanho da base de dados original, pois demoram apenas poucos minutos para serem executados, além de serem aplicados com pouca freqüência.

6 CONCLUSÕES

A velocidade de desenvolvimento dos sistemas de informações tem transformado alguns setores da sociedade. O setor elétrico, em especial, tem avançado continuamente no sentido da busca por eficiência e eficácia nos seus processos. Apontar causas de interrupções de energia é uma tarefa difícil para as equipes de restabelecimento de redes, e até secundário frente à necessidade do rápido restabelecimento, entretanto, o impacto dos apontamentos incorretos prejudica uma alocação adequada de investimentos e impacta no diagnóstico correto para a melhoria do desempenho dos indicadores de qualidade do sistema. Porém, observar os fatos correlacionados com cada interrupção e coletar as evidências associadas a um desligamento forçado, para uma posterior aplicação de técnicas como *KDD*, deve ser uma tarefa mais real para as equipes que trabalham em redes de distribuição. Dessa forma, a aplicação de *KDD* pode ser útil em banco de dados de empresas de distribuição, possibilitando a qualificação das informações utilizadas para planejar e operar os sistemas. Neste trabalho, buscou-se a construção de uma base de dados que representasse a correta identificação de causas de desligamentos forçados em sistemas de distribuição. Para alcançar esse objetivo, foram aplicadas técnicas que envolvem a manipulação e o tratamento de dados, bem como a seleção de regras que representassem o conhecimento que se tem desse tipo de sistema. Nesse caso, observou-se que a representação desse conhecimento através de simples regras de manipulação de dados possibilitou a qualificação das informações para o apoio ao planejamento e operação das redes de distribuição. A metodologia desenvolvida possui um ca-

ráter orientativo. A idéia base da metodologia proposta pode servir de guia para a qualificação de dados de desligamentos forçados nas empresas de distribuição de energia elétrica.

REFERÊNCIAS

Araribóia, G. (1987). *Inteligência Artificial: Um Curso Prático*. Livros Técnicos e Científicos Editora Ltda, Rio de Janeiro.

Billinton, R., R. Ghajar, F. Filippelli and R. Del Bianco (1995). The Canadian Electrical Association Approach to Transmission and Distribution Equipment Reliability Assessment. *Proc. of the IEE Conference on the Reliability of Transmission and Distribution Equipment*, Coventry, UK, pp. 7-12.

Brown, R.E. (2002). *Electric Power Distribution Reliability*. Marcel Decker, New York/Basel.

Buchanan, B.G. and E.H. Shortliffe (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, Massachusetts.

Chien, C-F., S-L. Chen and Y-S. Lin (2002). Using Bayesian Network for Fault Location on Distribution Feeder. *IEEE Transactions on Power Delivery*, Vol. 17, n^o 1, pp. 785-793.

Chow, M-Y., S.O. Yee and L.S. Taylor (1993). Recognizing Animal-Caused Faults in Power Distribution Systems Using Artificial Neural Networks. *IEEE Transactions on Power Delivery*, Vol. 8, n^o 3, pp. 1268-1274.

Congdon, P. (2003). *Applied Bayesian Modelling (Wiley Series in Probability and Statistics)*. First Edition, J. Wiley and Sons.

Costa, J. F. and Andrade Netto, M. L. (2007). Segmentação de mapas auto-organizáveis com espaço de saída 3-D. *Revista da SBA Controle & Automação*, Vol.18, n^o 2, pp.150-162.

Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA.

Han, J. and M. Kamber (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA.

Haykin, S. (2001). *Redes Neurais: princípios e prática*. Tradução: Paulo Martins Engel. Tradução de Neural Networks: a comprehensive foundation. 2nd ed. Bookman, Porto Alegre, RS.

Hsu, Y-Y., F-C. Lu, Y. Chien, J.P. Liu, J.T. Lin, P.H.S. Yu and R.R.T. Kuo (1991). An Expert System for Locating Distribution System Faults. *IEEE Transactions on Power Delivery*, Vol. 6, n^o 1, pp. 366-372.

MacQueen, J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, pp. 281-297.

Murphy, K.P. The Bayes Net Toolbox for Matlab. Department of Computer Science, University of California, Berkeley, CA, 2001.

Norvig, P. and S. Russel (2003). *Artificial Intelligence: a modern approach*. 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Oprisan, M., F. Filippelli, I. M. Clark and R. A. Billinton (1991). A Reliability data system for the reporting of forced outages of distribution equipment. *Proc. of the WESCANEX '91 'IEEE Western Canada Conference on Computer, Power and Communications Systems in a Rural Environment'*, Regina, Sask., Canada, pp. 267-270.

Peng, J-T, C.F. Chien, and T.L.B. Tseng (2004). Rough Set Theory for Data Mining for Fault Diagnosis on Distribution Feeder. *IEE Proceedings - Generation, Transmission, and Distribution*, Vol. 151, n^o 6, pp. 689-697.

Pretto, C.O., G.V. Rancich, F.A.B. Lemos and M.A. Rosa (2003). Forced Outages Information Treatment System and Cause Identification Based on Mobile Computing and Neural Networks. *Proc. of the IEEE Bologna Power Tech 2003 Conference*, Bologna, Italy, [em CD].

Pretto, C.O., F.A.B. Lemos and M.A. Rosa (2005a). Data Acquisition Using Mobile Computing Technology to Enhance Operation and Maintenance Planning. *Proc. of the 18th International Conference and Exhibition on Electricity Distribution (CIRED 2005)*, Turin, Italy, [em CD].

Pretto, C.O. (2005b). Sistema de Coleta e Tratamento de Informações sobre Desligamentos Não Programados Baseados em Computadores Móveis. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, PUC-RS, Porto Alegre, RS.

Pretto, C.O., M.A. Rosa, F.A.B. Lemos e T.T. Santos (2006). Utilização de computação móvel para qualificação de rotinas de operação e manutenção de redes de distribuição. *Revista da SBA Controle & Automação*, Vol. 17, n^o 4, pp. 446-458.

Teknomo, K. (2006). K-Means Clustering Tutorials. Disponível em: http://people.revoledu.com/kardi/tutorial/kMean/.

Tronchoni, A.B., C.O. Pretto, V. Licks, M.A. Rosa, F.A.B Lemos (2007). Forced Outage Cause Identification Based on Bayesian Networks. *Proc. of the 2007 IEEE Lausanne Powertech*, Lausanne, Switzerland, Vol. 1, pp. 1-6.

Xu, L., M-Y. Chow and L.S. Taylor (2007). Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm. *IEEE Transaction on Power Systems*, Vol. 22, n^o 1, pp. 164-171.

Xu, L. and M-Y. Chow (2006). A Classification Approach for Power Distribution System Fault Cause Identification. *IEEE Transaction on Power Systems*, Vol. 21, n^o 1, pp. 53-60.

Wehenkel, L. (1998). *Automatic Learning Techniques in Power Systems*. Kluwer