Análise de Inadimplência em Dados de Faturamento Utilizando Rede Bayesiana Ingênua Aumentada em Árvore

Cássio Dener Noronha Vinhal[1], Gélson da Cruz Jr[1], Luciana de Oliveira Berretta[1]

[1]Escola de Engenharia Elétrica e de Computação - UFG

cassio@eee.ufg.br, gcruz@eee.ufg.br, lberretta@nepe.eee.ufg.br

Resumo. *O presente trabalho verifica a aplicabilidade de Classificadores Bayesianos em Bancos de Dados de faturamento de uma distribuidora de energia. O intuito é encontrar padrões ou perfis em determinados grupos de consumo e estimar a quantidade de inadimplentes. O sistema computacional identifica padrões no histórico de cada cliente e projeta comportamentos prováveis. Utilizou-se o classificador Bayesiano Ingênuo Aumentado em Árvore em contraposição ao Bayesiano Ingênuo. A validação é feita através da comparação das taxas de acertos nas predições. As conclusões indicam uma abordagem adequada que oferece subsídios para se estabelecer políticas comerciais efetivas e reduzir a inadimplência.*

1 Introdução

A competitividade entre as empresas tem aumentado muito e para se manterem no mercado e obter uma maior lucratividade, essas empresas buscam novas e sofisticadas alternativas tecnológicas fundamentadas na matemática, sistemas de informação e engenharia para a redução de custos e aumento de lucros. Os resultados esperados permitem aumentar a variedade de produtos e serviços ofertados, conquistar novos mercados, realizar um "marketing" mais elaborado, adotar estratégias para manter o cliente e evitar a inadimplência.

O investimento em tecnologias permite observar que os processos administrativos estão se tornando cada vez mais informatizados e permitem o acúmulo de dados sobre compras e vendas, clientes, entre outros. Os dados permitem constituir Bancos de Dados que são armazenados em "Sistemas Gerenciadores de Bancos de Dados", compondo um grande histórico das transações das empresas e seus clientes. Entretanto dados produzidos e armazenados em larga escala muitas vezes não podem ser analisados por meio de métodos manuais tradicionais. Por outro lado, uma grande quantidade de dados pode vir a ser uma fonte de mais e melhores informações para a elaboração de políticas de negócios mais efetivas. Surge assim a necessidade de se explorar estes dados para extrair um conhecimento implícito, e. g., padrões ou regras importantes ali "escondidas" e que podem ser úteis para a tomada de decisões.

A utilização de técnicas como classificação, regras de associação, entre outras, tem aumentado muito. Nesse contexto o presente trabalho investiga modelos computacionais capazes de indicar a probabilidade de determinados consumidores se tornarem inadimplentes e fornecer medidas sobre a quantidade de inadimplentes, perfis de inadimplência dentro de grupos determinados, etc. Especificamente, um classificador

Bayesiano Ingênuo Aumentado em Árvore é analisado em contraposição ao Bayesiano Ingênuo e os resultados obtidos são muito promissores (Berretta, 2005).

2 Classificação

Classificação é uma tarefa muito importante para a identificação de padrões e predições. Em geral, uma classificação é uma função que permite determinar, a partir de um grupo pré-definido de rótulos, uma classe específica, segundo instâncias descritas por um conjunto de atributos.

Em Han et al. (2001), a classificação de dados é um processo onde, no primeiro passo, é construído um modelo que descreve um conjunto de classes predeterminadas ou conceitos. Este modelo é construído pela análise das amostras (tuplas) do Banco de Dados, descritas pelos atributos. As amostras individuais, escolhidas aleatoriamente, formam um conjunto de treino e procedimentos de aprendizado podem ser utilizados.

Classificadores Bayesianos são classificadores estatísticos. Eles podem predizer a probabilidade de um membro de uma classe e a probabilidade de uma dada amostra pertencer a uma classe particular.

3 Rede Bayesiana Ingênua

A Rede Bayesiana Ingênua (*Naïve-Bayes* (NB)) é uma estrutura simples com nós classificados como nós pais de todos os outros nós, sendo que nenhuma outra conexão é permitida.

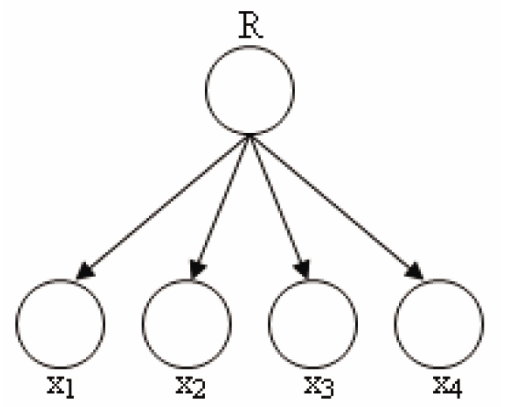

Figura 1 Exemplo de Rede Bayesiana Ingênua.

Segundo Mello (2001), a rede NB tem sido usada em classificadores há muitos anos, e possui duas vantagens sobre os outros classificadores: a) é de fácil construção e nenhum procedimento de aprendizagem é requerido b) o processo de classificação é muito eficiente computacionalmente, desde que ele assuma que todas as características são independentes das outras. Embora esta suposição possa ser problemática, a rede NB pode surpreender, apresentando-se superior a classificadores sofisticados onde as características não são fortemente combinadas.

O procedimento de construção de uma rede NB consiste basicamente em permitir que o nó de classificação seja o pai de todos os outros nós (nós filhos), não

sendo permitida a conexão entre os filhos. Não é necessário métodos para se levantar a estrutura da rede. Neste trabalho será utilizada uma rede Bayesiana diferente, cujos resultados obtidos nos estudos de caso serão comparados aos de uma NB. Portanto, a seguir é descrito um algoritmo para levantamento da estrutura de uma rede.

4 Algoritmo de Chow-Liu

O algoritmo utilizado para construir a estrutura da rede desse trabalho é baseado em um método conhecido na literatura como Algoritmo de Chow-Liu (1968) devido ao trabalho pioneiro cuja idéia é comparar distribuições diferentes sobre duas variáveis, consideradas dependentes ou independes, consistentemente com o domínio em que são estimadas, a partir de bancos de dados.

Um grafo não-direcionado é formado quando iniciado por um grafo sem arcos, adicionando-se um arco entre dois nós com máxima entropia. Logo após, um arco com máxima entropia associada é adicionado, desde que não crie um ciclo no grafo. Este processo é repetido até que não seja possível adicionar arcos. O passo final consiste em associar direções aos arcos de maneira a formar uma árvore. Pearl (1988) divide o método em duas fases. Na primeira fase ocorre a geração da árvore ponderada máxima, que produz um grafo não-direcionado contendo, a relação das variáveis do problema. Na segunda fase ocorre a definição da direcionalidade dos arcos.

A primeira fase é descrita na forma de algoritmo com cinco passos:

1)Dada uma distribuição P(x), computam-se as distribuições conjuntas $P(x_i, x_j)$ para todos os pares de variáveis;

2)Utilizando-se as distribuições calculadas no passo 1, calculam-se os pesos para todos os *n(n-1)/2* ramos da árvore, que devem ser ordenados por ordem de magnitude. Esses pesos são calculados pela equação da informação mútua, cujo desenvolvimento pode encontrado no trabalho de Pearl.

$$P(x_i, x_j) \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)}$$

3)Associam-se os dois ramos de maior peso a árvore a ser construída.

4)O próximo ramo da lista, já ordenada, deve ser acrescentado à árvore, contanto que não seja criado um ciclo. Caso isto aconteça, este ramo deve ser destacado e o próximo deve ser selecionado.

5)Repete-se o passo 4 até que *n-1* ramos tenham sido selecionados. Nesse ponto, o esqueleto da árvore está construído.

A segunda fase direciona os arcos, calculando a projeção de probabilidade de $P^{*}(x)$ sobre a distribuição P(x), selecionando um nó arbitrário para a raiz e formando o produto dado pela equação:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pais(x_i))$$

Sua complexidade é $O(n^2)$ e se utiliza apenas de comparações de pesos dos ramos.

5 Rede Bayesiana Ingênua Aumentada em Árvore

Uma Rede Bayesiana Ingênua Aumentada em Árvore (*Tree Augmented Naïve-Bayes* (TAN)) é uma estrutura que tem nós classificados como nós pais de todos os outros nós e permite conexões entre os nós filhos.

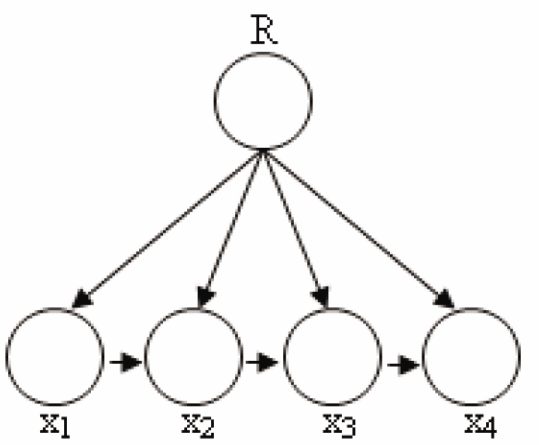

Figura 2 Exemplo de Rede Bayesiana Ingênua Aumentada em Árvore

Permitindo $X = \{x_1, \ldots, x_n, R\}$ representar o conjunto de nós (onde R é o nó de classificação) dos dados, o algoritmo para aprendizagem da classificação TAN aprende uma árvore estruturada sobre $X|\{R\}$, usando mútuos testes de informações. Ele, então, adiciona uma ligação do nó de classificação para cada característica do nó, à maneira como constrói uma rede Bayesiana Ingênua. Uma estrutura TAN simples é mostrada na Figura 2. Note que as características x_1, x_2, x_3 e x_4 formam uma árvore (CHENG et al., 2000).

Um procedimento de aprendizagem pode ser descrito como se segue:
1)Tomam-se o conjunto preparado e $X|\{R\}$ como entradas;
2)Chama-se o algoritmo Chow-Liu modificado e substitui-se todo o teste de informação mútua $I(x_i, x_j)$ por um teste de informação condicional $I(x_i, x_j)|\{R\}$;
3)Adiciona-se R como pai de todos x_i, onde $1 \leq i \leq n$;
4)Aprendem-se os parâmetros e produz-se a TAN.

A TAN pode ser descrita da seguinte forma:

$$\hat{c}_{TAN} = \arg\max_{c \in V} P(c) \prod_i P(a_i | Pais(a_i))$$

6 Modelo de Predição

Com o objetivo de realizar a predição de inadimplentes dos dados de faturamento de uma distribuidora de energia, foi desenvolvido um modelo que se utiliza de um classificador. Os classificadores podem predizer a probabilidade de um membro de uma classe e a probabilidade de uma dada amostra pertencer a uma classe particular.

Como elemento classificador foi utilizado uma Rede Bayesiana Ingênua Aumentada em Árvore (TAN). O algoritmo para aprendizagem da classificação TAN aprende uma árvore estruturada sobre X|{R} como já descrito, usando testes de informações mútuas.

Figura 3 Fases do Modelo

Com o intuito de facilitar a compreensão do modelo, este pode ser dividido em cinco etapas. A primeira etapa é o pré-processamento, responsável pela limpeza, desnormalização e discretização dos dados. A segunda etapa é a pré-seleção dos atributos, na qual a relevância dos atributos é analisada. A terceira etapa é o levantamento da estrutura da rede. A quarta é o levantamento das Tabelas de Probabilidade Condicional, que são as informações estatísticas dos dados. E finalmente, a fase de classificação, que faz a previsão dos possíveis inadimplentes.

6.1 Pré-Processamento

Na maioria das vezes, os dados não se encontram em um formato adequado, tornando-se necessário que os mesmos sejam tratados, a fim de permitir uma melhor aplicação do algoritmo de classificação. A etapa do Pré-Processamento é responsável por preparar esses dados para a análise.

De modo a promover a exatidão, a eficiência e a escalabilidade no processo de classificação, alguns passos devem ser seguidos, dentre os quais se destacam:

- Limpeza de Dados. Este procedimento, segundo Han et al. (2000), refere-se à remoção ou redução do “ruído” e ao tratamento dos dados ausentes. Isto pode ajudar a reduzir a confusão durante o aprendizado. Um aspecto importante desta fase é o tratamento dos valores ausentes;
- Desnormalização. O modelo de dados normalizado na 3FN (Terceira Forma Normal) pode requerer um maior número de junções para processar uma consulta e isso pode ser otimizado. A desnormalização retorna à 2FN (Segunda Forma Normal) ou à 1FN (Primeira Forma Normal), dependendo do caso;
- Discretização. A maioria dos algoritmos de aprendizagem de redes Bayesianas trabalha com variáveis categóricas (discretas não-ordenáveis), pois alguns

campos podem oferecer melhor desempenho na classificação se forem tratados como valores discretos. A técnica de discretização utilizada neste trabalho foi a dos K-Intervalos Proporcionais para Classificadores Bayesianos (Yang, 2003).

6.2 Pré-Seleção de Atributos

Na etapa de Pré-seleção dos Atributos é feita uma análise de relevância, pois muitos dos atributos dos dados podem ser irrelevantes para a tarefa de classificação. Além disso, outros atributos podem ser redundantes. Incluir atributos sem necessidade pode tornar o processo lento e possivelmente induzir a erro na etapa de treinamento.

6.3 Levantamento da Estrutura da Rede

Após as fases de pré-processamento e de pré-seleção dos atributos, faz-se necessário o levantamento da estrutura da rede. A estrutura da rede é uma representação abstrata do conhecimento do domínio, ou seja, a estrutura causal entre os processos do domínio.

Um modelo faz o levantamento da estrutura da rede de forma automática. O método utilizado para redes TAN já foi discutido (Seções 4 e 5). Ele toma uma distribuição de probabilidade P como entrada e constrói uma estrutura de árvore como saída. Para redes NB não é preciso levantar a estrutura, como explicado na Seção 3.

6.4 Levantamento da Tabela de Probabilidades Condicionais

Uma vez definida a estrutura da rede, é preciso especificar as probabilidades condicionais para os nós que participam diretamente das relações de dependência. Cada nó possui uma tabela de probabilidade condicional que quantifica a influência que os nós pais têm sobre cada nó filho. Essa construção consiste em levantar a probabilidade de cada nó X_i, dados seus pais (Pais(X_i)) - P(X_i|Pais(X_i)).

6.5 Classificação

As redes Bayesianas escolhidas para realizar o processo de classificação neste trabalho permitem encontrar probabilidades para todas as classes. A classe que apresentar a maior probabilidade é a escolhida como a classe do elemento.

7 Resultados

Os estudos de caso foram produzidos para grupos com predomínio de clientes da classe residencial (100% Residencial; 90% a 100% Residencial; 80% a 90% Residencial; 70% a 80% Residencial). Observou-se que em classes mais uniformes (e.g., 100% residencial) a redes NB e TAN apresentam desempenhos muito parecidos.

Cada grupo foi subdividido em bairros, com 5 faixas de consumo (abaixo de 100kWh/mês; 100kWh/mês a 200kWh/mês; 200kWh/mês a 300kWh/mês; 300kWh/mês a 500kWh/mês; 500kWh/mês a 800kWh/mês) como pode ser visto na Tabela 1. As primeiras faixas de consumo tendem a apresentar clientes de mais baixo consumo e também de menor poder aquisitivo.

Cada faixa direciona um experimento feito em 2 etapas, treinamento e classificação. Na etapa de treinamento, foram utilizadas amostras dos dados referentes aos meses de outubro, novembro e dezembro de 2002, e na etapa de teste, foram

utilizados dados dos meses de janeiro, fevereiro e março de 2003. Não foram utilizados dados do período de racionamento.

Em se tratando da predição de inadimplência, o erro foi medido em relação a valores históricos reais do Banco de Dados de uma distribuidora de energia. O desempenho das redes TAN é melhor que aquele das redes NB em todas as faixas do exemplo da Tabela 1, e.g., Bairro Jd. América. No desempenho global, considerando todos os estudos de caso, observa-se novamente que as redes TAN se sobressaem em relação às redes NB, conforme mostra a Tabela 2.

Tabela 1 Resultados - Bairro Jardim América

Faixa	Amostra	% Erro (NB)	% Erro (TAN)
1	12171	10,62	6,65
2	15587	2,09	1,90
3	7622	1,70	1,00
4	4948	1,01	0,76
5	1142	2,54	0,70

Tabela 2 Desempenho da TAN em relação à NB

	Ganhou	Perdeu	Empatou
Nº Casos	33	12	5
Percentual	66%	24%	10%

8 Conclusão

Após o desenvolvimento dos experimentos e analisando a abordagem proposta, pode-se concluir que, apesar da ocorrência de erros pequenos em sua maioria, os resultados são promissores. Para gerar resultados significativos, as amostras devem possuir um quantitativo mínimo para cada classe, a fim de evitar exclusividade na previsão de alguma classe.

O classificador TAN apresentou taxas de erro elevadas nas previsões realizadas para faixas de baixo consumo de energia (Tabela 1) e nas quais se verificaram taxas mais altas de inadimplência. Acredita-se que a escolha de atributos que possam refletir o perfil sócio-econômico em cada faixa, possa contribuir para uma diminuição da taxa de erros quando da caracterização de inadimplentes.

Os resultados também apontam para o desenvolvimento de novas políticas que definam inadimplência de uma forma mais flexível, permitindo a consideração de atrasos no pagamento por motivos diversos, flexibilidade de datas de pagamento, dentre outros, conforme apontado por especialistas da área de comercialização de energia.

Acredita-se que melhorias no modelo podem ser feitas em consideração a tais ponderações e que resultarão na melhoria das predições. Finalmente, a hipótese de o

classificador ser capaz de prever o comportamento dos consumidores pôde ser confirmada, e a TAN se mostrou mais precisa que a NB.

Referências

BERRETTA, L.O. (2005), "Análise de Inadimplência em Dados de Faturamento Utilizando Rede Bayesiana Ingênua Aumentada em Árvore", Dissertação de Mestrado, EEEC – UFG, Goiânia, Brasil.

CHOW, C., LIU, C. (1968), "Approximating Discrete Probability Distributions with Dependence Trees". IEEE Transactions on Information Theory, vol.14-3, 462- 467, USA.

CHENG, J.; GREINER, R. (2000) "Comparing Bayesian Network Classifiers". Alberta, CA: University of Alberta, 2000.

HAN, J.; KAMBER, M. (2000) "Data Mining, Concepts and Techniques". USA, Morgan Kaufmann,

MELLO, L. C. (2001) "Uma revisão de abordagens genético-difusas para descoberta de conhecimento em banco de dados". Porto Alegre, RS: Universidade Federal do Rio Grande do Sul - UFRS.

PEARL, J. (1988) "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference". USA, Morgan Kaufmann.

YANG, Y. (2003) "Discretization for Naïve-Bayes Learning". [S.l.]: School of Computer Science and Software Engineering of Monash University.