INSOLVENCY ANALYSIS ON ELECTRICITY BILLING DATABASES USING BAYESIAN CLASSIFIERS

GÉLSON CRUZ*, CÁSSIO VINHAL*, REINALDO NOGUEIRA*, EDUARDO REZENDE*,
LUCIANA BERRETTA*, FABRIZZIO SOARES*

*Núcleo de Estudos e Pesquisa em Energia, NEPE, Universidade Federal de Goiás
EEEC, Pça. Universitária s/n, Bloco A, Piso 3, 74605-220, Goiânia, GO
E-mails: gcruz@eee.ufg.br, cassio@eee.ufg.br, reinaldo@eee.ufg.br, eduardocr@yahoo.com.br, lucianaberretta@yahoo.com.br, fsoaresbr@yahoo.com.br

Abstract— The present work verifies the applicability of Bayesian Classifiers over Databases from an energy distribution company. The purpose is finding patterns or profiles into determined energy consumption groups and to estimate the number of insolvent clients. The predictive computational system identifies patterns related to each client historic and projects probable behaviors. Insolvency predictions from a Bayesian Network Augmented Naïve-Bayes (BAN) Classifier are compared to results obtained by a Tree Augmented Naïve-Bayes (TAN) Classifier and a Naïve-Bayes (NB) Classifier, taking into account historical insolvency records. Validity is verified by comparing prediction errors. Conclusions suggest an adequate approach which offers arguments for elaborating effective commercial policies for reducing insolvency.

Keywords— Electricity Billing, Bayesian classifiers, Insolvency prediction, Decision making

Resumo— O objetivo deste trabalho é fazer uma análise dos dados de faturamento de uma distribuidora de energia elétrica via classificadores Bayesianos. O propósito é encontrar padrões ou perfis em determinados grupos de consumo e estimar o número de clientes inadimplentes. O sistema automático de predição identifica padrões relacionados com o histórico de consumo dos clientes e projeta comportamentos prováveis. Predições de inadimplência de um classficador Bayesiano Ingênuo Aumentado em Redes Bayesianas (BAN) são comparadas àquelas obtidas por dois outros classificadores: o classificador Bayesiano Ingênuo (Naïve Bayes – NB) e o Bayesiano Ingênuo Aumentado em Árvore (TAN), considerando registros históricos de inadimplência dos consumidores. A validade é verificada pela comparação de erros de predição. As conclusões sugerem uma técnica adequada, a qual oferece argumentos para a elaboração de políticas comerciais para a redução de inadimplência.

Palavras-chave— Contas de Eletricidade, Classificadores Bayesianos, Predição de Insolvência, Tomada de Decisão

1 Introduction

Competition between electricity distribution companies has grown in the last decades. In order to stay in the market and to obtain a bigger profit, these companies search for new and sophisticated technological alternatives, based on mathematics, information systems and engineering solutions that can result on cost reductions and profit increases. Expected results allow decisions such as increasing the range of products and services offered, to conquer new markets, to make a more elaborated marketing, to adopt strategies for the customers and to avoid insolvency.

Investment on technologies allows us to observe that the administrative processes are becoming more and more computer-based and allow gathering data about sales, acquisitions, clients and much other relevant information. These data can be stored in Database Management Systems (DBMS), creating a huge historical of transactions of companies and their clients.

Although data produced and stored in large scale could not be analyzed by traditional hand-based methods, a large quantity of data can be used as a source of better and reliable information used to create more effective business policies. Also, a necessity of exploring these data in order to extract an implicit knowledge (e. g., "hidden" patterns or rules that can be useful for decision making) is created.

The utilization of techniques such as classification, association rules, etc, has increased in the last years. In this context this work analyses computational models capable of indicate the probability of some specific consumers become insolvent and give measures about the number of insolvents, their group profiles, etc. In order to accomplish this task, Bayesian Classifiers can be used with interesting results (Berretta, 2005), (Rezende, 2006).

2 Classification and Bayesian Networks

Classification is a very important task for the identification of patterns or predictions. Commonly, a classification is a function that allows determining, using a predefined group of labels, a specific class, according to instances described by a set of attributes.

In Han et al. (2000), data classification is a process where, during a first step, a model describing a set of predefined classes or concepts is built through the analysis of samples (tuples) of the database, described by their attributes. The randomly chosen individual samples form a training set and learning procedures can then be applied.

Bayesian classifiers are statistical classifiers. They can predict the probability of a class member and the probability of some predefined sample belonging to a particular class.

2.1 Naïve-Bayes Networks

The Naïve-Bayes Network (NB) is a simple structure with nodes classified as parent nodes of all other nodes, where no other connection is allowed.

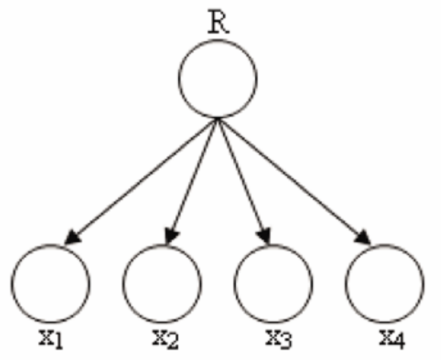

Figure 1. A Naïve-Bayes Network Example.

According to Mello (2001), NB networks have been used in classifiers for many years and have two advantages against others classifiers: a) They are of easy construction and no learning procedure is required b) the classification process is very computationally efficient, since it assumes that all the characteristics are independent of each other. Although this assumption could be a problem, the NB network can surprise the users, showing better results than sophisticated classifiers, where the characteristics are not strongly combined.

The building procedure of a NB network consists basically in allowing the classification node to be the parent of all other nodes (child nodes), while prohibiting the connection between child nodes. No method is necessary to build the net structure. In this work, results obtained using NB networks are compared to results obtained through different Bayesian networks.

2.2 The Chow-Liu Algorithm

The algorithm used to build the net structure in this work is based in a method known in the literature as the Chow-Liu algorithm (Chow et al., 1968), a pioneer work which main idea is to compare different distributions over two variables, considered dependent or independent, according with the domain where they are estimated, based on databases.

A non-directed graph is formed when, starting with a graph without arcs, an arc between two nodes is added with maximum entropy. After that, an arc with maximum entropy associated is added, since it does not create a cycle in the graph. This process is repeated until it is not possible to add more arcs. The final step consists in associating directions to the arcs in such a way to create a tree. Pearl (1998) divides the method in two phases. In the first phase the generation of the maximum weighted tree occurs, producing a non-directed graph containing the problem's variables relationship. In the second phase, the directionality definition of arcs occurs.

The first phase is described as an algorithm with five steps:

1) Given a distribution P(x), P(x*i*, x*j*) are computed as the joint distributions for all pairs of variables;
2) Using the distributions calculated in step 1, weights for all *n(n-1)/2* tree branches are calculated and must be ordered by magnitude order. These weights are calculated by the *mutual information equation*, whose development can be found in Pearl's work:

$$P(x_i,x_j)=\sum_{x_i,x_j}P(x_i,x_j)\log\frac{P(x_i,x_j)}{P(x_i)P(x_j)} \qquad (1)$$

3) The two branches related to the two bigger weights are associated to the tree being built;
4) The next branch of the list, already ordered, must be added to the tree, with the condition that a cycle is not created. If a cycle occurs, this branch must be discarded and the next on the list must be selected;
5) Step 4 is repeated until *n-1* branches are selected. At this point, the tree skeleton is built.

The second phase gives direction to arcs, calculating the probability projection of P'(x) over distribution P(x), selecting an arbitrary node for the root and forming the product given by the equation:

$$P'(x)=\prod_{i=1}^{n}P(x_i\mid Parents(x_i)) \qquad (2)$$

The complexity is O(n^2) and it only compares weights of branches.

2.3 Tree-Augmented Naïve-Bayes Network (TAN)

A Tree Augmented Naïve-Bayes (TAN) network is a structure with nodes classified as parent nodes of all other nodes and allows connections between the child nodes. Let X = {x_1,..., x_n, R} represent the set of data nodes (where R is the classifying node) of data, the TAN classification learning algorithm learns a structured tree over X|{R}, using mutual information tests and then adds a connection from the classification node to each node characteristic, as a Naïve-Bayes network is built. A simple TAN structure is shown in Fig. 2. It must be noticed that the characteristics x_1, x_2, x_3 e x_4 form a tree (Cheng et al., 2000).

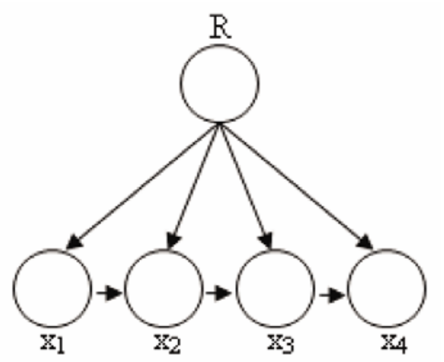

Figure 2. A Tree Augmented Naïve-Bayes Network.

A learning procedure can be described as follows:

1) A prepared set and X|{R} are taken as inputs;
2) The modified Chow-Liu algorithm is called and the mutual information test $I(x_i ,x_j)$ is replaced by a conditional information test $I(x_i , x_j)| \{R\}$);
3) R is added as parent of all x_i, where $1 \leq i \leq n$;
4) The parameters are learned and the TAN is produced.

The TAN can be described as:

$$v_{TAN} = \arg\max_{vj\in V} P(vj)\prod_{i} P(a_i \mid Parents(a_i)) \quad (3)$$

2.4 Tree-Augmented Naïve-Bayes Network

According to Mello (2001), this algorithm has the same basic idea of the TAN learning algorithm, but its second step calls an unconstrained bayesian network learning algorithm instead of Chow-Liu algorithm.

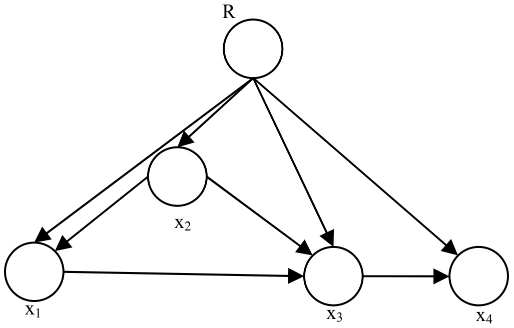

Figure 3. A Bayesian Network Augmented Naïve-Bayes.

Let X={x1,...xn,R} be the set of characteristics (where R is the classification vertex) of data, the learning procedure based on mutual information tests can be described as:

1) A prepared set and X\{R} (with the ordering) are taken as inputs;

2) Call the modified algorithm CBL1 (Cheng et al., 1997). The original algorithm is modified in this way:

Change all mutual information tests I(xi,xj) by conditional mutual information tests; change all conditional mutual information tests I(x_i, x_j, Z) by I(xi,xj |Z+{R}), where $Z\subset X\backslash\{R\}$. The conditional mutual information test equation is:

$$I(x_i,x_j \mid Z+\{R\})=$$

$$\sum_{x_i,x_j} P(x_i,x_j,Z+\{R\})\log\frac{P(x_i,x_j \mid Z+\{R\})}{P(x_i \mid Z+\{R\})P(x_j \mid Z+\{R\})} \quad (4)$$

3) Add R as parent of all x_i, where $1 \leq i \leq n$;
4) Learn the parameters and produces the BAN.

As the TAN learning algorithm, the BAN algorithm does not require any additional mutual information test and also requires $O(N^2)$ mutual information tests.

3 The Prediction Model

With the purpose of making a prediction of insolvents considering the billing data of a energy utility, it was developed a model that uses a classifier. Classifiers can predict the occurrence probability of a class member and the probability of some particular sample belonging to the class.

In order to facilitate comprehension of the model, five steps can be pointed out. The first step is the pre-processing, where data cleaning, de-normalization and discretization are made. The second step is the pre-selection of attributes, where attributes relevancy is analyzed. The third step consists on building up the network structure. The fourth step consists on building the Conditional Probability Tables which are data statistics. Finally, in the fifth step, the classification is performed and possible insolvents are predicted.

3.1 Pre-Processing

Most times data are not in adequate format, bringing the necessity of treating them in order to allow a better application of classification algorithm. The pre-processing step is responsible for preparing these data for analysis.

Some steps must be followed so as to promote exactness, efficiency and scalability of the classification process:

- Data Cleaning. This procedure, according to Han et al. (2000), refers to removal or reduction of “noise” and treatment of absent data. This may help to reduce confusion during the learning process;

- De-normalization. The data model normalized in the 3NF (Third Normal Form) can require a bigger number of joints to process a query, which can be optimized. De-normalization returns to the 2NF (Second

Normal Form) or to the 1NF (First Normal Form), depending on the case;

- Discretization. Most Bayesian network learning algorithms work with categorical variables (non-ordered and discrete), because some fields can offer a better classification performance if they are treated as discrete values. The discretization technique used here was the Proportional K-Intervals for Bayesian Classifiers (Yang, 2003).

3.2 Pre-Selection of Attributes

During this step, a relevancy analysis is made as many of data attributes can be irrelevant for the classifying task. Besides, other attributes can be redundant which would result in a slow process and possible errors during training process.

3.3 Building up the Network Structure

After pre-processing and pre-selection of attributes, it is necessary to build up the network structure. The network structure is an abstract representation of domain knowledge, i.e., the causal structure underlying domain processes.

A model builds up automatically the network structure. The method used for TAN networks was discussed earlier (Sections 4 and 5). It takes a probability distribution *P* as input and builds up a tree structure as output. For NB networks it is not necessary to build up the structure, as explained in Section 3. For BAN networks it is used the CBL1 algorithm (Cheng et al., 1997), that builds up the network analyzing relations of conditional independency among vertex and has a time complexity O(n^2). A probability distribution P is taken as input and the algorithm generates as output a BAN network.

3.4 Building the Conditional Probability Tables

Once defined the network structure, it is necessary to specify the conditional probabilities for the nodes that participate directly of dependency relations. Each node has a conditional probability table that quantifies the influence of parent nodes over each child node. This construction consists in finding each node X_i probability, given their parents ($Parents(X_i)$) - $P(X_i|Parents(X_i))$.

3.5 Classification

The Bayesian networks chosen for classification in this work, allow finding probabilities for all classes. The class with bigger probability is chosen as the element class.

4 Results

Case studies were carried out for groups with inherently residential clients (100% Residential; 90% a 100% Residential; 80% a 90% Residential; 70% a 80% Residential).

In order to perform the studies, each group was divided into districts, with 6 load demand levels (as shown in Table 1):

- below 100kWh/month;
- 100kWh/month to 200kWh/month;
- 200kWh/month to 300kWh/month;
- 300kWh/month to 500kWh/month;
- 500kWh/month to 800kWh/month;
- above 800kWh/month.

It was observed that in more uniform classes (e.g., 100% residential) the NB, TAN and BAN networks showed similar results. The results become different as the amount of residences decreases and the differences among the classifiers predictions are more evident. In this work, a result from a 70% Residential district was chosen to illustrate the insolvency prediction performance of each classifier.

The first load demand levels tends to aggregate clients with less electricity demand and also less buying power. Each level defines an experiment made in two phases: training and classification. During the training phase, data samples referring to October, November and December of 2002 were used and, during the test phase, data samples referring to January, February and March of 2003.

Considering the insolvency prediction, the error was calculated using historical values from a real energy utility database. The TAN networks performance is better than that obtained from NB networks considering all load demand levels, as indicated on Table 1 for a single district. Also, it was observed that BAN networks outperform TAN networks in most levels considered.

Table 1. Results – Jardim América District

Level	Sample	% Error (NB)	% Error (TAN)	% Error (BAN)
1	4898	11,35	6,78	5,58
2	5451	1,99	1,69	1,33
3	2766	1,52	0,71	0,80
4	1925	1,27	0,99	0,94
5	596	2,36	0,35	0,37
6	328	0,00	0,00	0,00

Tables 2 and 3 compare results obtained for BAN to those obtained for NB and TAN respectively. Those results were obtained through 72 cases, using 12 districts where the consumers are distributed into the 6 mentioned load levels.

Table 2. BAN versus NB relative performance

	Wins	Loses	Equals
Number of Cases	33	15	24
Percent (%)	45,83	20,83	33,33

Table 3. BAN versus TAN relative performance

	Wins	Loses	Equals
Number of Cases	29	10	33
Percent	40,28	13,89	45,83

5 Conclusion

After the development and analysis of experiments it can be concluded that even with the occurrence of small errors, results are promising. To generate significant results, data samples must include a minimum statistical quantitative for each class, to avoid exclusivity in prediction of some class.

All classifiers showed high error rates in predictions performed for low load demand levels (Table 1) where high insolvency was verified also. It is believed that a better choice of attributes in order to reflect social-economical profile of each level may contribute for decreasing the error rates when characterizing insolvents.

Results also points out for the development of new policies that define insolvency in a more flexible way, allowing the utilization of different criteria for characterizing payment delays and to study the impact of flexible payment dates, as observed by energy commercialization specialists.

Acknowledgments

The authors would to thank CELG – Companhia Energética de Goiás for financial and technical support on this research.

References

Berretta, L. O. (2005). Análise de Inadimplência em Dados de Faturamento Utilizando Rede Bayesiana Ingênua Aumentada em Árvore, *MSc. Dissertation*, EEEC-UFG, Goiânia, GO, Brazil.

Cheng, J., Bell, D. A. and Liy, W. (1997). An algorithm for Bayesian belief network construction from data, *Proceedings of AI & STAT'97*, Florida, USA, pp. 83-90.

Cheng, J.; Greiner, R. (2000). Comparing Bayesian Network Classifiers. Alberta, CA, University of Alberta.

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

Han, J.; Kamber, M.(2000). Data Mining, Concepts and Techniques. Morgan Kaufmann, USA.

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

Pearl, J. (1998). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, USA.

Rezende, E.C. (2006). Análise de Inadimplência em Dados de Faturamento Utilizando Rede Bayesiana Ingênua Aumentada em Redes Bayesianas. *MSc. Dissertation*, EEEC – UFG, Goiânia, GO, Brazil.

Yang, Y. (2003). Discretization for Naïve-Bayes Learning. School of Computer Science and Software Engineering of Monash University.