Knowledge discovery for CRM improvement

G. M. Caputo, V. M. Bastos, A. M. Cister & N. F. F. Ebecken
COPPE – Federal University of Rio de Janeiro, Brazil

Abstract

The objective of this paper is to present a database marketing analysis through data and text mining tools. A case study of a Brazilian Power Energy distribution was developed indoors. The main idea is to transform the database information into strategic marketing knowledge. Thus a data warehouse sample was treated, reduced and clustered. Principal component analysis was used to reduce the original number of variables. The entire database was classified after creation by the decision trees and neural networks approach. In this work, text mining techniques were used to process customers' claims in order to improve cluster results. The CRM group has developed a powerful tool to gather knowledge regarding the skills and habits of customers, thereby gaining their confidence and loyalty.

Keywords: data mining, text mining, clustering, CRM.

1 Introduction

This work shows a methodology to extract knowledge of customer habits from an electric energy company database, in order to discover useful information that will serve as base for CRM implementation in the company.

For this, segmentation variables have been used that identify partner-economic characteristics, behaviors and customer invoicing data, which represent the customer profile for the company.

The set of initial variables extracted from the company Data Warehouse involved a great number of dimensions, with many forms of customers' segment identification. Through the data extraction process, it was possible to get samples with good universal representation, allowing the application of knowledge extraction techniques, in order to identify customers' behavior.

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080171

First, some mathematical transformations had been made on the samples, making it possible to eliminate missing values and those whose magnitude is bigger than the average existing values (outliers).

After cleaning and transforming the data, it was possible to process the information, through the data mining tools application, presenting significant results for customers' behavior knowledge.

From the use these results obtained in the data knowledge extraction phase, it was possible to obtain impact information that allows actions to be directed towards problem.

Thus, diffuse and little structured information, which is difficult to interpret due to its form as it is presented, was converted into clear and simple knowledge to be easily understood.

2 Measures for customer satisfaction evaluation

The database classification process is carried out through the execution of two stages that include original data conversion to the format processed by the data mining solution proposal and a data mining algorithm application on the converted base.

2.1 Data conversion

The data used for the customer satisfaction evaluation had been obtained from the company Data Warehouse, which has historical data of 2 million customers regarding consumption, invoicing, insolvency and attendance services amongst other information. From the entire database 5600 registers were extracted corresponding to the active customers between 2003 and 2006.

From the extracted registers two groups were been created, called Group A and Group B. Group A is composed of high energy consumers whose consumption is above 350Kw/month, while Group B represents the majority of the company's customers, whose consumption is below 350Kw/month.

The treatment made on each values set from each variable of the sample is presented in the following procedure consisting of six steps.

1st step: Data in secular format
Each register of the Data Warehouse's table stores only one invoice for each customer, as shown in Table 1. V_j represents the variables used in the grouping, where j=1… n, and n is the number of analyzed database attributes, and m=*jan*,…,*dec*, represents the months for each analyzed year.

After that, it was necessary to transpose the table, as shown in Table 2. This implies that each customer would have only one register in the table, and all the invoices are present in this register, where each attribute is represented as a month/attribute pair.

2º step: Variable average of the same month among evaluated years
From the data given in Table 2, the averages of every month were calculated, with the objective of reducing the number of variables to be analyzed. The averages are made in expression (1), shown below.

$$V_{1\,jan}|_{2003} + V_{1\,jan}|_{2004} + V_{1\,jan}|_{2005} + V_{1\,jan}|_{2006} = (\sum \text{jan} \mid_{03\text{-}06}) / 4 \qquad (1)$$

where $V_{1\,jan}|_{2003}$ represents the "attribute #1" value of January 2003 and $(\sum \text{jan} \mid_{03\text{-}06}) / 4$ represents the "attribute #1" average values of months January 2003 to January 2006.

Table 1: Data format in invoicing information.

Customers	V_1	V_2	…	V_n
C_1	$V_{1\,jan}$	$V_{2\,jan}$		$V_{n\,jan}$
C_1	$V_{1\,feb}$	$V_{2\,feb}$		$V_{n\,feb}$
C_1	$V_{1\,mar}$	$V_{2\,mar}$		$V_{n\,mar}$
…		…		…

Table 2: Temporal analysis.

Customers	V_1 - 2003			…	V_1 - 2006		
C_1	$V_{1\,jan}$	…	$V_{1\,dec}$	…	$V_{1\,jan}$	…	$V_{1\,dec}$
C_2	$V_{1\,jan}$	…	$V_{1\,dec}$	…	$V_{1\,jan}$	…	$V_{1\,dec}$
…	…	…	…	…	…	…	…

3º step: Adding the averages
The annual averages generated in the previous step are added in order to obtain only one value for each variable, as shown in equation (2).

$$(\sum \text{jan} \mid_{03\text{-}06}) / 4 + (\sum \text{feb} \mid_{03\text{-}06}) / 4 + \ldots + (\sum \text{dec} \mid_{03\text{-}06}) / 4 \qquad (2)$$

4º step: Weighed mean. Identification of each month value contribution.
This mean is obtained applying equation (3).

$$\frac{\sum jan \mid_{03-06}}{\sum_{i=1}^{12} \sum i \mid_{03-06}} = C_{jan}\,\% \qquad (3)$$

5º step: Multiplication of the averages contribution (smoothing) – eliminating the seasonality.

To eliminate the influences caused by seasonality equation (4) is applied on the value obtained in the previous step.

$$\sum_{jan} \Big|_{03-06} * C_{jan} \% \tag{4}$$

6º step: Adding the12 means – variable analysis
This is calculated using equation (5).

$$\sum_{i=1}^{12} \sum i \Big|_{03-06} * C_i \% \tag{5}$$

2.2 Artificial neural network

2.2.1 Group A results

With a neural network it is possible to identify a function with fewer classification errors. Thus, it becomes easy to develop an automatic procedure that allows classification of all database customers with other information sources, such as customer claims.

Table 3: Confusion matrix for group A.

	Cluster	**1**	**2**	**3**		**Cluster**	**1**	**2**	**3**
	1	229	1	4		**1**	233	0	2
	2	1	14	0		**2**	1	15	0
RNA1	**3**	6	0	28	**RNA2**	**3**	2	0	30
	Cluster	**1**	**2**	**3**		**Cluster**	**1**	**2**	**3**
	1	233	2	0		**1**	231	0	1
	2	1	13	0		**2**	2	15	0
RNA3	**3**	2	0	32	**RNA4**	**3**	3	0	31
	Cluster	**1**	**2**	**3**					
	1	**233**	0	1					
	2	1	**15**	0					
RNA5	**3**	2	0	**32**					

To execute the neural network algorithm 284 records were extracted from the sample. The confusion matrix shows cases that data were classified correctly and incorrectly, according to the information given previously. Table 3 shows that RNA5 is the neural network with the best behavior and Table 4 represents neural networks performance and summarizes the confusion matrix information.

Table 4: Neural network performance on group A.

	RNA1			RNA2			RNA3		
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Total	236	15	32	236	15	32	236	15	32
Corrects	229	14	28	233	15	30	233	15	32
Errors	7	1	4	3	0	2	0	0	0
Corrects (%)	93,33	97,03	87,5	98,72	100	93,7	86,66	98,72	100
Errors (%)	6,667	2,967	12,5	1,27	40	6,25	13,33	1,27	0
	RNA4			**RNA5**					
Cluster	**1**	**1**	**1**	**1**	**2**	**3**			
Total	236	15	32	236	15	32			
Corrects	231	15	31	233	15	32			
Errors	5	0	1	3	0	0			
Corrects (%)	97,88	100	96,87	98,72	100	100			
Errors (%)	2,11	0	1,27	1,27	0	0			

The sensitivity analysis presented in Table 5 shows the confidence degree of the results in uncertain classifications or assumptions about data and results used. Therefore, it is essential to investigate differences between the data. This analysis is similar to the communality analysis that verifies the importance of each variable and its contribution in the proposed model. In this case, the variable that most contributes to the neural network RNA5 is invoiced demand outside the typical amount.

2.2.2 Group B results

Group B is composed of 4181 registers extracted from the sample. Another 119 registers were chosen randomly to validate the results, in order to test the function created from the best performing neural network.

The confusion matrix presented in Table 6 shows cases where data were classified correctly and incorrectly, according to the previous classification.

Table 7 shows the confusion matrix information given by the neural networks performance and Table 8 identifies RNB4 as the neural network ranks that best represent the right classification.

Table 5: Sensitivity analysis of group A.

	Price	Consumption quantity	Power factor hour tip quantity	Invoiced demand outside tip amount	Contracted demand outside tip amount	Contracted demand out HUMID tip quantity
RNA1						
Ratio	1,99	1,43	2,05	1,46	1,2	1,36
Importance	2	4	1	3	6	5
RNA2						
Ratio	1,09	1,3	0,89	1,34	1,02	1,02
Importance	3	2	6	1	5	4
RNA3						
Ratio	2,94	1,85	2,38	2,86	1,18	1,2
Importance	1	4	3	2	6	5
RNA4						
Ratio	1,05	1,34	1,07	1,02	0,99	0,97
Importance	3	1	2	4	5	6
RNA5						
Ratio	2,42	1,38	1,74	2,71	1,38	1,40
Importance	2	5	3	1	6	4

Table 6: Confusion matrix for group B.

	Cluster	**1**	**2**	**3**	**4**
RNB1	**1**	133	0	0	2
	2	0	3	0	0
	3	0	0	3903	0
	4	0	2	0	19

	Cluster	**1**	**2**	**3**	**4**
RNB2	**1**	132	0	1	2
	2	0	3	0	1
	3	1	2	3092	0
	4	0	0	0	18

	Cluster	**1**	**2**	**3**	**4**
RNB3	**1**	133	0	0	2
	2	0	5	0	0
	3	0	0	3093	0
	4	0	0	0	19

	Cluster	**1**	**2**	**3**	**4**
RNB4	**1**	**133**	0	0	0
	2	0	**5**	0	0
	3	0	0	**3093**	0
	4	0	0	0	**21**

	Cluster	**1**	**2**	**3**	**4**
RNB5	**1**	132	0	1	0
	2	0	5	0	0
	3	0	0	3092	0
	4	1	0	0	21

Table 7: Neural network performance on group B.

	RNA1				RNA2				RNA3			
Cluster	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Total	133	5	3903	21	133	5	3903	21	133	5	3903	21
Corrects	133	3	3903	19	132	3	3902	18	133	5	3903	19
Errors	0	2	0	2	1	2	1	3	0	0	0	2
Corrects (%)	100	60	100	90,5	99,3	60	100	85,7	100	100	100	90,5
Errors (%)	0	40	0	9,25	0,75	40	0,03	14,3	0	0	0	9,52
	RNA4				**RNA5**							
Cluster	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**				
Total	133	5	3903	21	133	5	3903	21				
Corrects	133	5	3903	21	132	5	3902	21				
Errors	0	0	0	0	1	0	1	0				
Corrects (%)	100	100	100	100	99,3	100	100	100				
Errors (%)	0	0	0	0	0,75	0	0,03	0				

Table 8: Sensitivity analysis of group B.

	Consumption	Mean consumption in a month	Supply cost	Service Rate	Spare cut value	Total bill value
RNB1						
Ratio	7,43	6,8	8,95	7,37	5,96	7,92
Importance	3	5	1	4	6	2
RNB2						
Ratio	2,29	2,12	2,07	2,11	2,74	1,64
Importance	2	3	5	4	1	6
RNB3						
Ratio	4,1	6,11	7,63	7,78	5,14	6,12
Importance	6	4	2	1	5	3
RNB4						
Ratio	10,64	6,08	9,21	12,89	8,25	7,94
Importance	**2**	**6**	**3**	**1**	**4**	**5**
RNB5						
Ratio	5,8	6,82	11,03	10,58	9,1	17,32
Importance	6	5	2	3	4	1

The sensitivity analyses presented in Table 8 show the results confidence degree in situations of uncertain judgments or assumptions about the data and results used. Therefore, it is essential to investigate differences between data, variables and algorithms, since the importance of each variable contributes to the proposed model. In this case, the variable that most contributes to the RNB4 neural network is "Service Rate".

Table 9 shows that a test of 119 records confirms the neural networks error percentage. Therefore, the best function to classify the entire database is generated by the **RNB4** neural network.

Finally, a cross analysis with these results and some categorical variables was executed, in order to identify customer cluster characteristics in both groups. Such categorical variables identify customers' behavior such as geographic localization, connection type (residential, industrial, etc.), invoicing, type of measurer, etc. Thus, the conclusions about obtained clusters are shown in Table 10.

Table 9: Neural network errors percentage.

	Neural Network				
	RNB1	RNB2	RNB3	RNB4	RNB5
Error Percentage	0.098%	0.172%	0.049%	0.000%	0.049%

Table 10: Cross analysis with customer characteristics.

	Cluster	Customer	%	Invoicing	Geography	Connection	Measurer
Group A	1	Medium	83,39	Medium to low	Oceânica	Residential and commercial	Not identified
	2	Heavy Users	5,3	High	Guanabara, Serrana	Commercial and industrial	Not identified
	3	Big Customers	11,31	Medium to high	Oceânica, Norte	Commercial and Services	Not identified
Group B	Cluster	Customer	%	Invoicing	Geography	Connection	Measurer
	1	Intermediate	3,56	Medium	Oceânica, Norte	Residential	Three-phase
	2	Heavy Users	0,19	High	Oceânica, Serrana	Not residential	Three-phase
	3	Below Zero	95,74	Low	Guanabara	Residential	Mono-phase
	4	Medium	0,51	Medium to high	Oceânica	Not residential	Three-phase

2.3 Clusters and claims

After the identification of customer profiles, it was possible to execute a cross analysis with clusters and categorical variables that represent claims registered in the attendance for group B customers. These variables identify types of claims classified in the call center.

The graph shown in figure 1 presents the characteristics of each cluster in accordance with customers' claims [1]. Note that many claims are concentrated in Customers Attendance, Attendance Quality, Quality Supply and Invoicing.

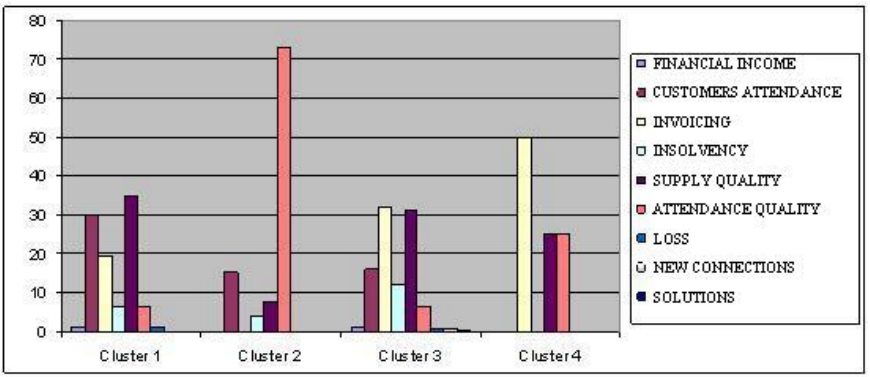

Figure 1: Customers' claims and clusters.

3 Conclusions

The methodology described in this paper, when applied to a company's data, can mean a sufficiently useful tool to identify customer satisfaction measures. These measures can be obtained from a database sample, where pre-selected attributes, such as invoicing, insolvency and attendance services information, identify customers' behavior.

This tool is based on the execution of some activities through the data, described as follow:

- Sample Data Generation – consists of selecting a significant subgroup of data that expresses the customers' behavior;
- Cleaning and Conversion Data – missing values and outliers' removal, followed by data conversion from the existing format in the DW to the format for the data mining algorithm;
- Data Mining Algorithm Execution –converted data is processed for posterior results analysis;
- Complete Database Reclassification – consists of segmentation identification of all customers.

All analysis on large data sets, except in censuses, is made on representative universe samples [2]. To classify the entire database is extremely complex, since the functions generated from the sample easily are not applied to all the data. In

such a way, this tool helps analysts to achieve information classification, and consequently in the CRM execution [3].

Many data mining techniques allow classification of the entire database from samples, independently of the universal size. In the present work the neural network technique was used, which is easily applicable to high complexity problems.

The development of a filtering tool [4] made it possible for the user to extract knowledge of the entire database, in order to identify useful information related to the company's customers. Moreover, it is possible to combine the results with other kinds of information, such as customers' claims, in order to complement the knowledge.

With this methodology being applied by the company, the results can, in fact, be used in the identification of customers' behavior evaluation. The distinct treatment of each customer segment directs the adequate marketing techniques application, consequently providing an increase in customer satisfaction.

CRM is one of the most important marketing actions that directly influence company success. Thus, it is necessary to emphasize its use for intelligence group managers with respect to union and planning of CRM-enterprise vision.

Acknowledgements

We are grateful to the Brazilian Research Agencies CNPq and FAPERJ for their financial support.

References

[1] Caputo, G. M., Bastos, V. M., Ebecken, N. F. F. – Data Mining VII: Using text mining to understand the call center customers' claims – WIT Press - Ashurst Lodge – U.K. 2006
[2] Kish, L, Survey Sampling, John Wiley & Sons, Canada, 1965.
[3] Cister, A. M., Ebecken, N. F. F. – Data Mining III: CRM through DM: a case study – WIT Press – Ashurst Lodge – U.K. 2002.
[4] Cister, A. M., Ebecken, N. F. F. – Data Mining IV: A screening tool based on neural networks – WIT Press – Ashurst Lodge – U.K. 2004.