Data Mining Applied to CRM
for a Long Distance Telephone Company

HUGO L.C. AZEVEDO, MARLEY M.B.R. VELLASCO, EMMANUEL P.L. PASSOS
Laboratório de Inteligência Computacional Aplicada
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de S. Vicente, 225, Gávea,
Rio de Janeiro, RJ, Brasil, 22453-900
BRASIL

Abstract: - Nowadays, it is very important for retail companies to understand their customers and establish a good relationship with them. It is then crucial to be able to segment the customers according to their buying patterns and needs. Unfortunately, this is not an easy task, requiring the consideration of several different mathematical techniques. In this work, a database from a long distance telephone company containing client-calling patterns was clustered and characterized. First, the Kohonen algorithm was used to cluster a subset of the database. Then, the clusters extracted were characterized using rules and some statistical visualization techniques. Finally, a classification model, based on neural networks, was built to classify future customers into the clusters extracted.

Key-Words: - data mining, KDD, neural networks, Kohonen Algorithm, clustering, pattern classification, CRM

1 Introduction

Nowadays, it is very important for a company to understand its customers and to create and manage relationships with them (CRM- Customer Relationship Management) [7]. In markets where the competition is high, this becomes a survival issue. Therefore, it is very important for a company to be able to segment its clients into clusters with similar characteristics. Segmentation allows the company to provide specific services and products to each group, according to its needs. A good performance in this task, usually increases the level of satisfaction and fidelity of the clients, consequently increasing company's revenues and profit. Such task is usually easy for companies with few clients. For large companies, however, the situation is very different, requiring sophisticated techniques to accomplish an efficient segmentation. In such companies, client databases are generally large and complex, with usually spread all over the company departments.

Clustering, cluster characterization and classification, are some of the tasks that are usually performed in a process called data mining, which is the most important step of a larger process called KDD (Knowledge Discovery in Databases) [3]. The mining step usually uses statistics, computational intelligence, OLAP, and other techniques, with the ultimate goal of extracting some useful and non trivial knowledge from a complex database.

In the present work, a KDD process was performed on a complex database containing calling patterns from the clients of a long distance telephone company.

2 Description of the Work

2.1 Database

The database analyzed in this work contained summarized data about the last 3 monthly bills of a sample of 4,000 company's clients. The study was limited to non-international long-distance calls made from/to fix (not mobile) phones. The data retrieved from the company's data warehouse was condensed into a single table containing 49 attributes. The attributes were related to:

- Distance between the caller and the phone called;
- Day of the week in which the call was made (weekday, Saturday or Sunday);
- Time period of the call (peak or off-peak);
- Type of the call (inter-regional, intra-regional or intra-sector).

The measures used to quantify the attributes above were:

- Number of calls made in a month;
- Number of minutes spent in a month;
- Revenues generated by the client in a month.

For each measure, the average of the last 3 months was calculated.

2.2 Knowledge Discovery

A KDD process includes the following steps: 1- Data Cleaning; 2 – Data Selection; 3 – Data Codification; 4 – Data Mining and 5 – Validation and Interpretation of the Results. All these steps were performed in this work. It's worth saying that the whole KDD process is very iterative, since there can be many loops on the path from step 1 to step 5, as well as very interactive. This means that many steps were repeated several times, until good results were obtained.

In step 1, some consistency checks were made on the database in order to fix or eliminate some inconsistent data.

In step 2, after a few analyses, the most significant client attributes were selected. The selected attributes are shown in Table 1. In addition, many different subsets were selected from database. The different subsets were used to train, test and validate the models built on step 4. Care was taken such that each subset (dataset) would be representative of the whole database.

In step 3, following suggestions from [8], different types of codification were tested. In the end, it was decided to use the simpler one, where all attributes were normalized to [0,1].

ATTRIBUTES	DESCRIPTION
RWDAY	Revenues from calls made on weekdays
RSATURDAY	Revenues from calls made on Saturdays
RSUNDAY	Revenues from calls made on Sundays
RINTER	Revenues from inter-regional calls
RINTRAR	Revenues from intra-regional calls
RINTRAS	Revenues from intra-sector calls
RPEAK	Revenues from calls made during peak time
ROPEAK	Revenues from calls made during off-peak time
MINUTES	Consumption in minutes in a month
REVENUES	Total revenues in a month

Table 1 – Descriptions of the attributes chosen. Each attribute contains the average value of the last 3 months.

As can be noticed, steps 1, 2 and 3 are performed only to prepare the data to the data mining step, which actually takes place in step 4. Step 4, the most important one, was split into three phases: clustering; characterization of the clusters; and building of a classification model. In the clustering phase, the clients were split into groups with similar characteristics using the *Kohonen* algorithm [6]. On the second phase, the main characteristics of each cluster were extracted using classification rules and statistical visualization techniques [5]. Finally, in the last phase, a classification model was built to classify future clients. Sections 3, 4 and 5 describe in detail those 3 phases.

In step 5, the results from step 4 were analyzed and validated. The results and analyses are presented in the end of sections 3, 4 and 5 and in section 6.

3 Clustering

Over 20 attempts were carried out trying to split clients into clusters. In each attempt, different parameters for the *Kohonen* algorithm and different datasets were used. Each of these attempts was called a study. To choose the parameters, the suggestions presented [4] and [6] were followed. This section contains the studies which produced the best results. The clustering process was carried out on Matlab 5.3. After each run of the algorithm, the result was analyzed with the help of two two-dimensional density maps (*Kohonen* Maps). The first map, contains over each neuron on the grid a number representing the number of patterns associated with that neuron. The second map, contains the same information represented by circles of different sizes describing different ranges of values.

Before the clustering identification process, it was analyzed, in a general way, how patterns were arranged throughout the map, according to their main characteristics. It was found, for example, that patterns associated with clients generating low revenues were mainly concentrated on the upper left corner of the map, while patterns associated with clients in the opposite situation were mainly concentrated in the opposite corner. This and other observations were taken into consideration when identifying the clusters. Care was also taken to avoid creating clusters that would correspond to a small or very large percentage of the whole dataset. Finally, as can be verified in the next section, its worth saying that the whole process is very subjective, depending strongly on the analyst's criteria to find the clusters.

3.1 Studies

In the first study (S1) it was used a 20X20 map, 2000 and 200.000 iterations on phases I and II, respectively, a learning rate decay from 0.1 to 0.02 on phase I and a neighborhood equal to zero on phase II.

Fig. 1 contains the resultant map found in S1. In this map, there seems to be 4 groups of patterns, but only the group on the lower right corner seems to have a reasonably clear border around it. Therefore, such group was assumed as the first cluster found (Fig. 2). In order to make the task of finding more clusters easier, all the patterns associated with the cluster just found were removed from the dataset. The new dataset was again submitted to the *Kohonen* algorithm, and after a few attempts obtained the results of study S2 were obtained.

In the second study (S2) it was used a 18X18 map, 2000 and 160.000 iterations on phases I and II, respectively, a learning rate decay from 0.4 to 0.02

on phase I and a neighborhood equal to 1 on phase II.

Fig. 3 contains the map found in study S2. In this map, again, there seems to be 4 groups of patterns, but only the group on the upper left corner seems to have a reasonably clear border around it. This group was then assumed as a cluster and all the patterns associated with it were removed from the dataset. This new dataset was again submitted to the *Kohonen* algorithm in a new study S3. This process was repeated twice more on studies S4 and S5, and the results for this last study are presented next.

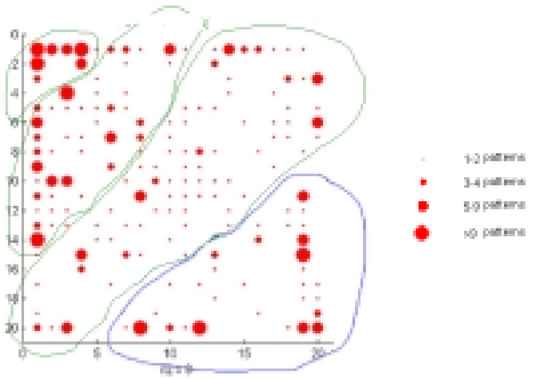

Fig. 1 – *Kohonen* Map for S1 with possible clusters circled.

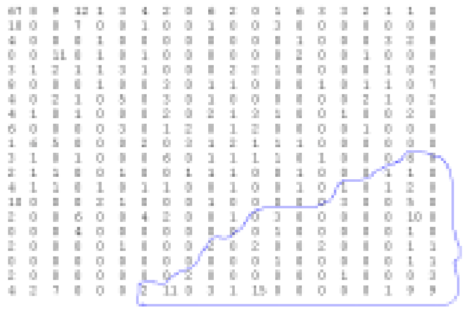

Fig. 2 - *Kohonen* Map for S1 with the cluster found circled.

In study S5 it was used a 12X12 map, 1000 and 70.000 iterations on phases I and II, respectively, a learning rate decay from 0.1 to 0.02 on phase I and a neighborhood equal to 1 on phase II.

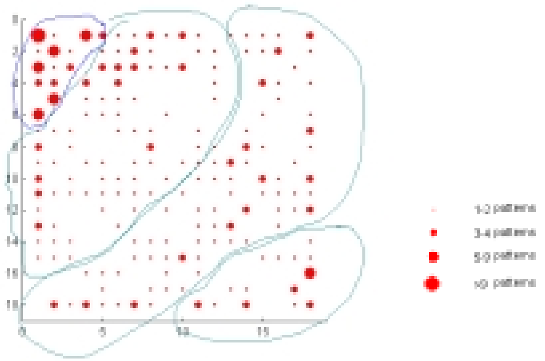

Fig. 3 - *Kohonen* Map for S2 with possible clusters circled.

In this last study (Fig. 4), it was possible to visualize 3 distinct groups of patterns with well-defined borders around them. These 3 groups were then assumed as clusters.

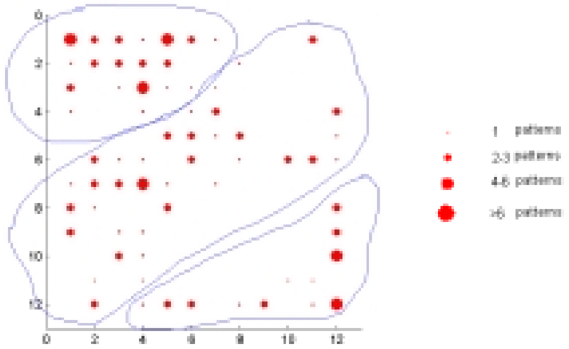

Fig. 4 - *Kohonen* Map for S5 with clusters found circled.

3.2 Results

After study S5, seven clusters had been established – one in each of the studies S1 to S4 and three in S5. However, after a few analyses in the characterization phase, it was verified that the cluster found in S1 could, and should, be re-divided into other clusters. In the characterization phase, it was possible to see that such cluster had three smaller groups inside it, each with characteristics a little different from one another. These 3 groups can also be seen in Fig. 1, one to the right and below, another to the left and below and the last one above and to the right

The whole clustering process produced 9 clusters. The percentages of each cluster in the dataset used are presented in Table 2. In order to visualize and validate the clustering process, the original dataset from S1 was again submitted to the *Kohonen* algorithm, but this time coloring the patterns according to the clusters they belong (Fig. 5).

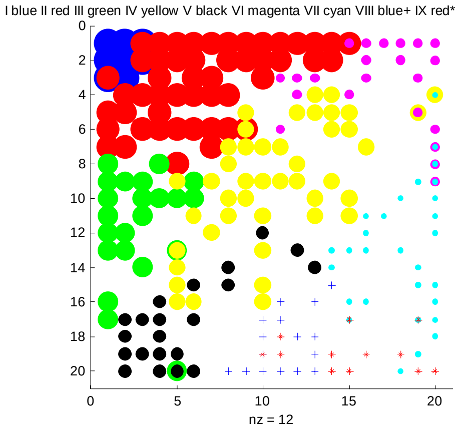

Fig. 5 – Clusters found in the clustering phase.

Even using just one single network to classify the whole dataset into all the 9 clusters, it can be verified from Fig. 5 that there is little overlap among clusters, showing that the clustering phase was able to segment the dataset in different groups.

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	Total
% of the total	20,0	22,6	7,8	13,6	6,0	9,4	9,2	6,2	5,2	100

Table 2 – Percentages of each cluster in the dataset.

4 Clusters Characterization

In this phase of the mining process, the main characteristics of the patterns of each cluster were extracted and analyzed. To do so, it was used the software *WizRule* [9], which is based on combinatory optimization, to generate rules describing the clusters and the statistics package SPSS to generate some graphs and descriptive statistics. First, the graphs were analyzed [5] and some hypotheses about the profile of the clusters formulated. Then, other graphs and the rules generated by *WizRule* were used to corroborate those hypotheses. This process led to the results shown in section 4.3.

The analyses were based on suggestions from the Marketing Dept. of the company who provided the database. The analyses were the following:

- Level of revenues generated (amount of money spent monthly by the client);
- Level of consumption in minutes (total number of minutes spent monthly by the client);
- Revenues on weekdays X revenues on weekends;
- Revenues on long distance calls X revenues on short distance calls;
- Revenues on different periods.

Section 4.1 presents some of the histograms, box plots and scatter plots generated with the aid of SPSS. Section 4.2 presents some of the rules generated by *WizRule*. Finally, section 4.3 presents the results from the characterization phase.

4.1 Graphs in SPSS

The histogram in Fig.6 gives an idea of how the revenues are distributed among clients and what should be considered as low revenues, intermediary level revenues and high revenues. In the box plot shown in Fig. 7, it can be seen that, for example, revenues generated by clients from clusters III, V and VIII are relatively higher on weekends, while revenues generated by clients from clusters II, IV, VI and VII are relatively higher on weekdays.

Analyzing the scatter plot in Fig. 8, it can also be seen that revenues from clients from cluster VI are more concentrated on calls during peak-time.

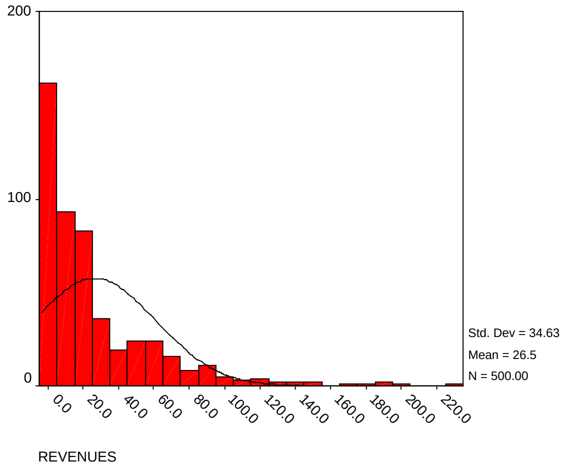

Fig. 6 – Histogram of the variable REVENUES.

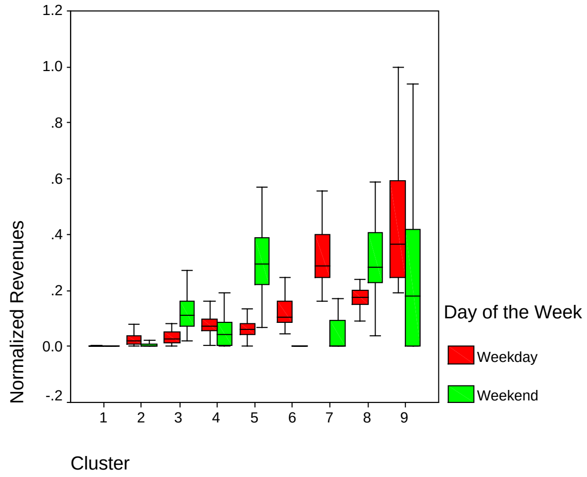

Fig. 7 – Box plot comparing weekday X weekend revenues.

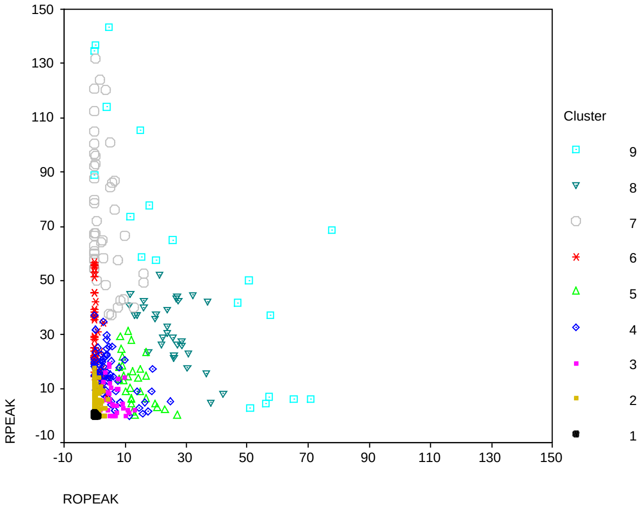

Fig. 8 – Scatter plot comparing peak X off-peak revenues.

4.2 Rules Generated by *WizRule*

Many attempts (studies) were carried out trying to find rules to describe the clusters. In each attempt, different confidence and support levels were used [1]. For a rule of the type *If A then B*, the support level is defined as the ratio between the number of rows where the rule holds and the total number of rows. The confidence level is defined as the ratio between the number of rows where the rule holds and the number of rows where *A* holds. Fig. 9 presents a reduced subset of the rules generated by *WizRule* in the two main studies. In one of them, a support level of 10% and a confidence level of 70% were used. With these parameters, it was possible to find rules describing some of the characteristics of clusters I, II, VIII and IX. In the other study, a support level of 4% and a confidence level of 50% were used. In this last study, it was possible to find rules for the remaining clusters.

```
3)     If REVENUES is 0.05 ... 1.52 (average = 0.60 )
       Then   CLUSTER is 1.00

78)    If REVENUES is 1.58 ... 3.60 (average = 2.72 )
       Then   CLUSTER is 2.00

93)    If RWDAY is 1.12 ... 2.99 (average = 2.00 )
        and RSUNDAY is 0.00 ... 0.09 (average = 0.01 )
       Then   CLUSTER is 2.00

346)   If RWDAY is 0.00 ... 1.03(average = 0.32)
        and RINTER is 1.94 ... 3.17(average = 2.68)
       Then   CLUSTER is 3.00

390)   If ROPEAK is 7.05 ... 7.59(average = 7.34)
       and RINTRAR is 0.00
       Then   CLUSTER is 3.00

373)   If RWDAY is 17.06 ... 18.68(average = 18.04)
        and RINTRAS is 0.00
       Then   CLUSTER is 4.00

379)   If RINTER is 18.26 ... 18.89(average = 18.48)
        and REVENUES is 18.26 ... 19.57(average = 18.76)
       Then   CLUSTER is 4.00

391)   If  RSUNDAY is 8.03 ... 12.94(average = 10.73)
        and RPEAK is 0.00 ... 3.48(average = 1.07)
       Then   CLUSTER is 5.00

393)   If RINTRAR is 0.00
        and RINTRAS is 0.00
        and MINUTES is 205.00 ... 214.00(average = 209.17)
       Then   CLUSTER is 5.00

279)   If RSATURDAY is 0.00
        and RSUNDAY is 0.00
        and MINUTES is 78.00 ... 87.00(average = 83.33)
       Then   CLUSTER is 6.00

433)   If RWDAY is 18.98 ... 20.59(average = 19.97)
        and ROPEAK is 0.00 ... 0.09(average = 0.01)
       Then   CLUSTER is 6.00

333)   If RPEAK is 83.62 ... 197.41(average = 129.09)
        and RINTRAR is 0.00 ... 0.15(average = 0.02)
       Then   CLUSTER is 7.00

77)    If ROPEAK is 19.81 ... 66.80 (average = 34.09 )
       Then   CLUSTER is 8.00 or 9.00

85)    If RSATURDAY is 7.00 ... 33.00 (average = 11.35 )
       Then   CLUSTER is 8.00 or 9.00
```

Fig. 9 – Some of the rules generated by *WizRule*.

The rules generated by *WizRule* led to conclusions less straightforward but similar to those obtained before, with SPSS. For example, rules 279 and 433 agreed to the suppositions made for cluster VI (relatively higher revenues on weekdays), earlier on section 4.2. The same was true for rule 93 and the suppositions made for cluster II.

4.3 Results

Based on the rules generated by *WizRule* and on the graphs from SPSS, a profile for each cluster was created. The profiles found are shown next:

Cluster I: very low revenues, no defined pattern for the calls.
Cluster II: low revenues, more characterized by calls made on weekdays.
Cluster III: low intermediary level revenues, more characterized by inter-regional calls made on weekends during off-peak time.
Cluster IV: intermediary level revenues, more characterized by inter-regional calls made on weekdays during peak time.
Cluster V: intermediary level revenues, more characterized by long inter-regional calls made on weekends during off-peak time.
Cluster VI: intermediary level revenues, more characterized by short intra-sector calls made on weekdays during peak time.
Cluster VII: high revenues, more characterized by inter-regional calls made on weekdays on peak time.
Cluster VIII: high revenues, more characterized by long inter-regional calls made on weekends during peak and off-peak time, specially on the latter.
Cluster IX: very high revenues, more characterized by inter-regional calls made in all days of the week.

5 Classification Model

In the last phase of the mining process, a classifying model was developed to classify future clients and the clients who were left out of the dataset clustered in the clustering phase. Instead of using the *Kohonen* network as the classification model, it was decided to create and verify the performance of two other models, one based on classification rules and another based on the *BackPropagation* neural network. *Aspentech NeuralSIM* (formerly known as *NeuralWorks Predict*) was used to build the model based on neural networks. *WizWhy* [9] was used to build the model based on rules. After building both models, their performance was compared and the best chosen.

5.1 Classification with *NeuralSIM*

NeuralSIM is a software tool based on neural networks that, before training and running the network, pre-process the training data and tries to find the best network architecture to suit the performance needs. It uses *BackPropagation* to train the network and provides possibility to choose many of its parameters. Since it is based on neural networks, it does not provide explanations (e.g. classification rules) for the classification process.

Three attempts (studies) were conducted to build a classification model with *NeuralSIM*. In each study, several nets were tested on different datasets. Study S0 used mostly the default parameters of the software. Each network was trained with the adaptative gradient learning algorithm. Some of the main parameters used in each study, as well as the best network architecture found by the software, are shown in Table 3.

Parameter	**S0**	**S1**	**S2**
Hidden Layer Function	Tanh	Tanh	Sigmoid
Output Layer Function	Softmax	Softmax	Softmax
Network Evaluation Function	Avg Classific. Rate	Avg. Classific. Rate	Accuracy
Nets Trained	1	3	4
Architecture Found	16-4-9	16-8-9	14-15-9

Table 3 – Main parameters and architectures used in each study.

Among the models generated by *NeuralSIM*, S1 was chosen because it produced the most homogeneous performance combined with the second highest accuracy rate (Table 4).

Model / Cluster	**S0 (NeuralSIM)**		**S1 (NeuralSIM)**		**S2 (NeuralSIM)**	
Type of Error	A	B	A	B	A	B
I	100%	96,8%	100%	96,8%	96,7%	96,7%
II	91,2%	100%	94,1%	100%	97,0%	97,1%
III	91,7%	64,7%	83,3%	71,4%	91,7%	84,6%
IV	66,7%	100%	76,2%	100%	90,5%	100%
V	100%	75%	100%	69,2%	100%	81,8%
VI	85,7%	80%	85,7%	92,3%	92,9%	100%
VII	85,7%	85,7%	100%	87,5%	100%	77,8%
VIII	100%	100%	100%	100%	100%	90%
IX	87,5%	87,5%	87,5%	100%	37,5%	100%
Total	89,4%	89,4%	92,1%	92,1%	92,7%	92,7%

Table 4 – Accuracy considering false positives (A) and false negatives (B).

5.2 Classification with *WizWhy*

WizWhy is a software tool that, besides generating classification rules to describe a database, also makes classifications/predictions based on the rules discovered. As *WizRule*, it is also based on combinatory optimization and works in a similar way. It has the advantage of generating rules to justify the predictions accomplished. The software also provides the probability that the prediction is true. Unfortunately, the algorithm implemented in the software is able to make only Boolean predictions. To overcome such problem, 9 sets of rules, describing each of the 9 clusters, were generated. Then, for each client in the validation dataset, 9 predictions were made, and the probability of each prediction being true was stored. Finally, in a procedure similar to a Bayesian Classifier [2], the 9 probabilities were compared and the client was classified into the cluster with the higher probability associated.

5.3 Results

To evaluate both models, their performance considering false positives and false negatives was verified and compared. The average performance of the model built with *NeuralSIM* (92,1%, for study S1) was considerably higher than the performance of the model built with *WizWhy* (72,3%), which led the former to be chosen as the classification model for the work. The relative bad performance of the model built with *WizWhy*, was, perhaps, in part due to the difficulty in choosing, in a homogenous way, the support and confidence parameters for all of the 9 sets of rules. In other words, for each set of rules, different support and confidence parameters had to be chosen according to the data, which may have biased the predictions.

6 Final Comments

Despite the complexity of the data, the results obtained were satisfactory and useful for the telephone company. Since different techniques were used to accomplish the same task, it was also possible to obtain a general idea of the advantages and disadvantages of the techniques used.

It is intended, for a future project, to solve other case studies assigning different weights to the attributes, according to their importance, and trying to integrate the *Kohonen* algorithm with visualization techniques. It is also intended to use neuro-fuzzy systems in the classification task to conjugate the advantages of neural networks with the capability of rule extraction from fuzzy systems.

References

[1] Agrawal, R., Imielinski, T., Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings, ACM SIGMOD Conference on Management of Data, Washington, D.C.

[2] Duda, R. Hart, P., Pattern Classification and Scene Analysis, John Wiley & Sons, 1973

[3] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, MIT Press, 1996

[4] Haykin, S., Neural Networks – A Comprehensive Foundation, Prentice Hall, 2nd edition, 1999

[5] Johnson, R. Wichern, D., Applied Multivariate Statistical Analysis, Prentice Hall, 4th edition, 1999

[6] Kohonen, T., Self-Organizing Maps, Springer, 2nd edition, 1997

[7] Mckenna, R., Relationship Marketing, Perseus Publishing. Campus, 1993

[8] Weiss, S. M., Indurkhya, N., Predictive Data Mining. Morgan Kaufmann Publishers, Inc, 1998

[9] WizRule, WizWhy, www.wizsoft.com