Data Mining for Decision Support in e-banking area

Vasilis Aggelis
University of Patras
Department of Computer Engineering and Informatics
Rio, Patras, Greece
Vasilis.Aggelis@egnatiabank.gr

Dimitris Christodoulakis
University of Patras
Department of Computer Engineering and Informatics
Rio, Patras, Greece
dxri@cti.gr

ABSTRACT

The introduction of data mining methods in the banking area due to the nature and sensitivity of bank data, can already be considered of great assistance to banks as to prediction, forecasting and decision support. Concerning decision making, it is very important a bank to have the knowledge of (a) customer profitability and their grouping according to this parameter and (b) association rules between products and services it offers in order to more sufficiently support its decisions. Object of this paper is to demonstrate that keeping track of customer groups according to their profitability and discovery of association rules between products and services it offers to those groups, is of major importance as to its decision support.

Keywords: Data Mining, Decision Support, Association Rules, Clustering, RFM analysis.

1. Introduction

RFM analysis [DataPlus Millenium, (2001)] is a three-dimensional way of classifying, or ranking, customers to determine the top 20%, or best, customers. It is based on the 80/20 principle that 20% of customers bring in 80% of revenue. It is of high significance concerning decision support.

In order to group customers and perform analysis, a customer segmentation model known as the pyramid model [Curry, J. and Curry, A. (2000)] is used. The pyramid model groups customers by the revenue they generate, into the categories shown in Figure 1. These categories or value segments are then used in a variety of analytics. The advantage of this approach is that it focuses the analytics on categories and terminology that are immediately meaningful to the business.

The pyramid model has been proven extremely useful to companies, financial organisations and banks. Indicatively some issues that can be improved by the use of the model follow:

- Decision support and Decision making.
- Future revenue forecast.
- Customer profitability.
- Predictions concerning the alteration of customers' position in the pyramid.
- Understanding the reasons of these alterations.
- Conservation of the most important customers.
- Stimulation of inactive customers.

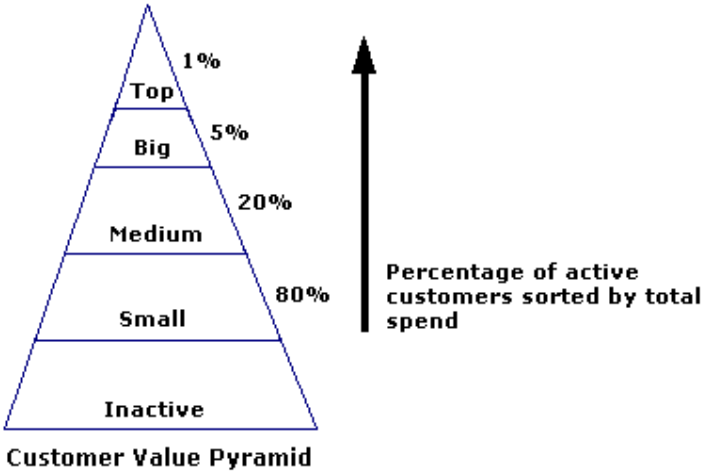

Figure 1

Essentially RFM analysis suggests that the customer exhibiting high RFM score should normally conduct more transactions and result in higher profit for the bank.

RFM analysis [SPSS (2001), Madeira, S.A. (2002), COMPAQ (2001), Im, K. and Park, S. (1999)] nowadays can be conducted by the use of Data Mining methods like clustering. These methods contribute to the more efficient determination and exploitation of RFM analysis results.

Determination of association rules concerning bank data is a challenging though demanding task since:

- The volume of bank data is enormous. Therefore the data should be adequately prepared by the data miner before the final step of the application of the method.
- The objective must be clearly set from the beginning. In many cases not having a clear aim results in erroneous or no results at all.
- Good knowledge of the data is a prerequisite not only for the data miner but also for the final analyst (manager). Otherwise wrong results will be produced by the data miner and unreliable conclusions will be drawn by the manager.
- Not all rules are of interest. The analyst should recognize powerful rules as to decision making.

The challenge of the whole process rises from the fact that the relations established are not easily observed without use of data mining methods.

The increase of electronic transactions during the last years is quite rapid. E-banking nowadays offers a complete sum of products and services facilitating not only the retail customers but also corporate customers to conduct their transactions easily and securely. The ability to discover rules between different electronic services is therefore of great significance to a bank.

Identification of such rules offers advantages as the following:

- Description and establishment of the relationships between different types of electronic transactions.
- The electronic services become more easily familiar to the public since specific groups of customers are approached, that use specific payment manners.
- Customer approach is well designed with higher possibility of successful engagement.
- The improvement of already offered bank services is classified as to those used more frequently.

- Reconsidering of the usefulness of products exhibiting little or no contribution to the rules.

Of certain interest is the certification of the obtained rules using other methods of data mining to assure their accuracy.

In the present paper, the RFM scoring of active e-banking users is studied along with the ranking of these users according to the pyramid model. This study is also concerned with the identification of rules between several different ways of payment in each customer group. The software used is SPSS Clementine 7.0. Description of various clustering techniques and algorithms as well as association rules' basic features follow in section 2 while in section 3 the calculation of the RFM scoring of active e-banking users and the process of investigation for association rules is described. Section 4 contains experimental results derived from the data sets of section 3 and finally section 5 contains the main conclusions of this work accompanied with future work suggestions in this area.

2. Clustering and Association Rules basics

2.1. Clustering Techniques

Clustering techniques [Hand, D, Mannila, H., Smyth P. (2001), Collier, K., Carey, B., Grusy, E., Marjaniemi, C., and Sautter, D. (1998)] fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables.

Clustering techniques are used for combining observed examples into clusters (groups) that satisfy two main criteria:

- each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
- each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

Depending on the clustering technique, clusters can be expressed in different ways:

- identified clusters may be exclusive, so that any example belongs to only one cluster.
- they may be overlapping; an example may belong to several clusters.
- they may be probabilistic, whereby an example belongs to each cluster with a certain probability.
- clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

2.2. K-means algorithm

K-means [Hand, D, Mannila, H., Smyth P. (2001), Collier, K., Carey, B., Grusy, E., Marjaniemi, C., and Sautter, D. (1998), Bradley, P. and Fayyad, U. (1998), Zha, H., Ding, C., Gu, M., He, X. and Simon, H. (2001)] is the simplest clustering algorithm. This algorithm uses as input a predefined number of clusters that is the *k* from its name. Mean stands for an average, an average location of all the members of a particular cluster. When dealing with clustering techniques, a notion of a high dimensional space must be adopted, or space in which orthogonal dimensions are all attributes from the table of analysed data. The value of each attribute of an example represents a distance of the example from the origin along the attribute axes. Of course, in order to use this geometry efficiently, the values in the data set must all be numeric and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space.

K-means algorithm is a simple, iterative procedure, in which a crucial concept is the one of *centroid*. *Centroid* is an artificial point in the space of records that represents an average location of the particular cluster. The coordinates of this point are averages of attribute values

of all examples that belong to the cluster. The steps of the K-means algorithm are given below.

1. Select randomly *k* points (it can be also examples) to be the seeds for the *centroids* of *k* clusters.
2. Assign each example to the *centroid* closest to the example, forming in this way *k* exclusive clusters of examples.
3. Calculate new *centroids* of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (*centroid*).
4. Check if the cluster *centroids* have changed their "coordinates". If yes, start again form the step 2). If not, cluster detection is finished and all examples have their cluster memberships defined.

Usually this iterative procedure of redefining *centroids* and reassigning the examples to clusters needs only a few iterations to converge.

2.3. Association rules

A rule consists of a left-hand side proposition (antecedent) and a right-hand side (consequent) [Hand, D, Mannila, H., Smyth P. (2001)]. Both sides consist of Boolean statements. The rule states that if the left-hand side is true, then the right-hand is also true. A probabilistic rule modifies this definition so that the right-hand side is true with probability p, given that the left-hand side is true.

A formal definition of association rule [Han, E., Karypis, G., and Kumar, V. (1997), Brin S., Motwani, R., and Silverstein, C. (1997), Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2002), Michail A. (2000), Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T.(1996), Srikant, R. and Agrawal, R. (1995), Toivonen, H. (1996)] is given below.
Definition. An association rule is a rule in the form of

$$X \rightarrow Y$$

Where *X* and *Y* are predicates or set of items.

As the number of produced associations might be huge, and not all the discovered associations are meaningful, two probability measures, called *support* and *confidence*, are introduced to discard the less frequent associations in the database. The support is the joint probability to find X and Y in the same group; the confidence is the conditional probability to find in a group Y having found X.

Formal definitions of support and confidence [Zaki, M.J., Parthasarathy, S., Li, W., and Ogihara M. (1997), Hand, D, Mannila, H., Smyth P. (2001), Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2002), Michail A. (2000), Srikant, R. and Agrawal, R. (1995), Toivonen, H. (1996)] are given below.
Definition Given an itemset pattern X, its *frequency fr(X)* is the number of cases in the data that satisfy X. *Support* is the frequency fr($X \wedge Y$). *Confidence* is the fraction of rows that satisfy Y among those rows that satisfy X,

$$c(X \rightarrow Y) = \frac{fr(X \wedge Y)}{fr(X)}$$

In terms of conditional probability notation, the empirical accuracy of an association rule can be viewed as a maximum likelihood (frequency-based) estimate of the conditional probability that Y is true, given that X is true.

2.4. Apriori algorithm

Association rules are among the most popular representations for local patterns in data mining. Apriori algorithm [Zaki, M.J., Parthasarathy, S., Li, W., and Ogihara M. (1997), Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1996), Hand, D, Mannila, H., Smyth P. (2001), Han, E., Karypis, G., and Kumar, V. (1997), Ng, R., Lakshmanan, L. and Han. J. (1998), Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2002)] is one of the earliest for finding association rules. This algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules. This algorithm contains a number of passes over the database. During pass k, the algorithm finds the set of frequent itemsets L_k of length k that satisfy the minimum support requirement. The algorithm terminates when L_k is empty. A pruning step eliminates any candidate, which has a smaller subset. The pseudo code for Apriori Algorithm is following:

C_k: candidate itemset of size k
L_k: frequent itemset of size k

L_1 = {frequent items};
For (k=1; L_k != null; k++) do begin
    C_{k+1} = candidates generated from L_k;
    For each transaction t in database do
        Increment the count of all candidates in
        C_{k+1} that are contained in t
    L_{k+1} = candidates in $C_{k+1}$with min_support
    End
Return L_k;

3. RFM scoring of active e-banking users

The data sample used concern the period between January 1st and December 12th of the year 2002.
The term «active e-banking user» describes the user who has conducted at least one financial transaction during this period. In order RFM scoring to express customer profitability, all values concerning financial transactions are taken into consideration.

The following variables are calculated for this specific time period.
Recency (R)
R is the date of the user's last transaction. Since the R value contributes to the RFM scoring determination, a numeric value is necessary. Therefore, a new variable, R_{new} is defined as the number of days between the first date concerned (1/1/2002) and the date of the last active user's transaction. For example a user who has conducted his last transaction on 29/11/2002 is characterized by R_{new}=332, while one who has conducted his last transaction on 4/4/2002 will have R_{new}=93.
Frequency (F)
R is defined as the count of financial transactions the user conducted within the period of interest (1/1/2002 to 12/12/2002).
Monetary (M)
M is the total value of financial transactions the user made within the above stated period.

RFM Score (RFM Factor) is calculated using the formula:

$$RFM\_Factor = R_{new}+F+M.$$

In order to have a normalized sum of the above variables, we divide the total value of financial transactions with 1,000. All the variables contribute the same to the RFM_Factor result.

A sample of the data set on which data mining methods are applied lies in Table 1.

User	R_{new}	F	M (*1,000 €)	RFM_Factor
...	...	...	...	...
User522	330	20	€208,56	558,56
User523	216	6	€169,32	391,32
User524	304	8	€128,66	440,66
User525	92	1	€272,45	365,45
...	...	...	...	...

Table 1

The sample includes 1904 active users in total. Customer classification is performed using the K-means method.

In order to search for relations between different payment types of the customers of each group, in this study the following payment types were used.

1. *SWIFT Payment Orders* – Funds Transfers among national and foreign banks.
2. *Funds Transfers* – Funds Transfers inside the bank.
3. *Forward Funds Transfers* – Funds Transfers inside the bank with forward value date.
4. *Greek Telecommunications Organization (GTO) and Public Power Corporation (PPC) Standing Orders* – Standing Orders for paying GTO and PPC bills.
5. *Social Insurance Institute (SII) Payment Orders* – Payment Orders for paying employers' contributions.
6. *VAT Payment Orders.*
7. *Credit Card Payment Orders.*

For the specific case of this study the medium e-banking customers, both individuals and companies, were used. A Boolean value is assigned to each different payment type depending on whether the payment has been conducted by the users or not.

A sample of the above data set is shown in Table 2.

User	SWIFT P.O.	FUNDS TRANSFER	FORWARD FUNDS TRANSFER	GTO-PPC S.O.	SII P.O.	VAT P.O.	CREDIT CARD P.O.
…	…	…	…	…	…	…	…
User103	F	T	T	F	F	F	T
User104	T	T	T	F	T	T	F
User105	F	T	T	T	T	T	T
User106	F	T	F	T	F	F	T
…	…	…	…	…	…	…	…

Table 2

In total the sample contains 298 medium e-banking customers.

In order to discover association rules the Apriori [SPSS (2002)] method was used. These rules are statements in the form *if antecedent then consequent*.

4. Experimental results

As seen in the histogram of Figure 2, RFM distribution is high over values less than 1.000.This is a natural trend since, as concluded in paragraph 1, 80% of the customer exhibits low RFM Factor.

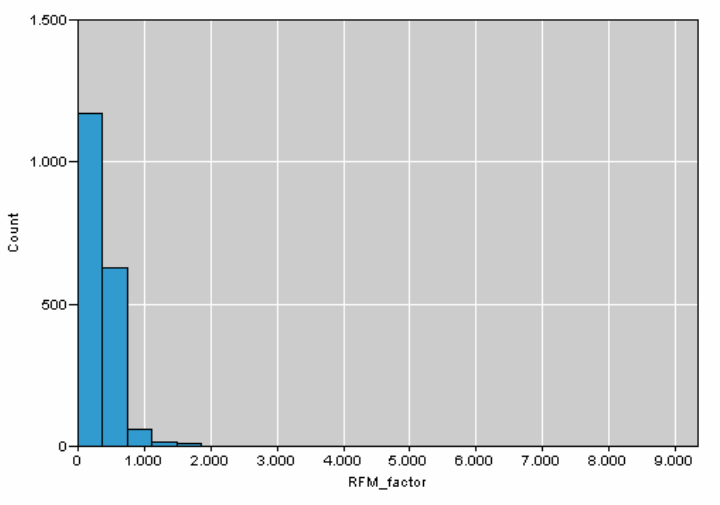

Figure 2

Application of the K-means algorithm results in the 4 clusters of Figure 3. Next to each cluster one can see the number of appearances as well as the average value of each variable.

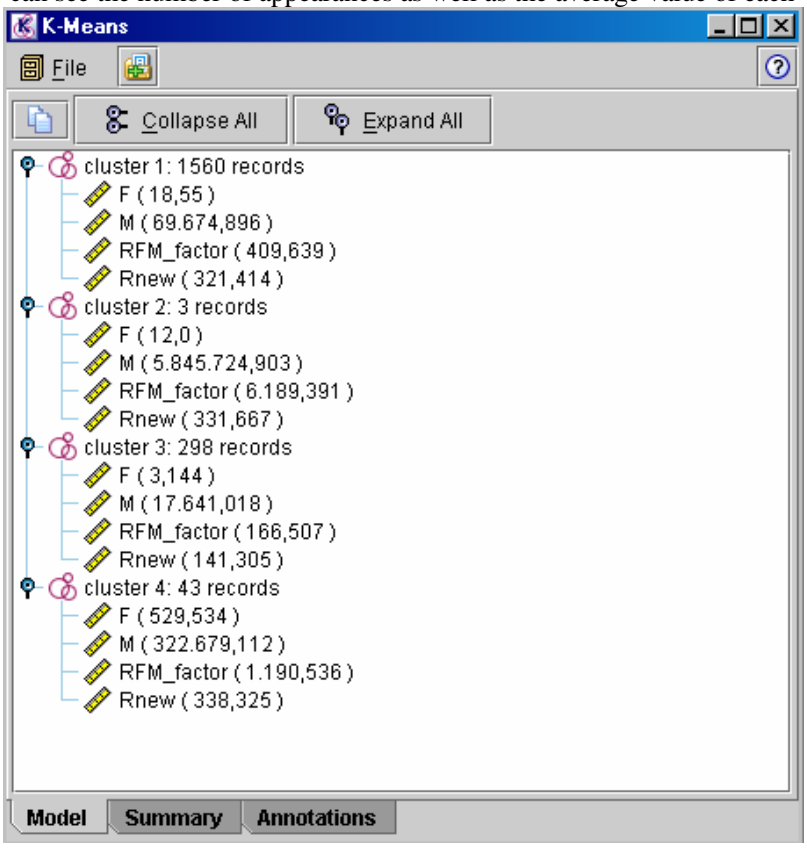

Figure 3

The above clustering results in the distribution of Figure 4. The similarity of this distribution to the pyramid model is apparent:

Cluster 1	(81,93%)	⇨	Small	80%
Cluster 3	(15,65%)	⇨	Medium	15%
Cluster 4	(2,26%)	⇨	Big	4%
Cluster 2	(0,16%)	⇨	Top	1%

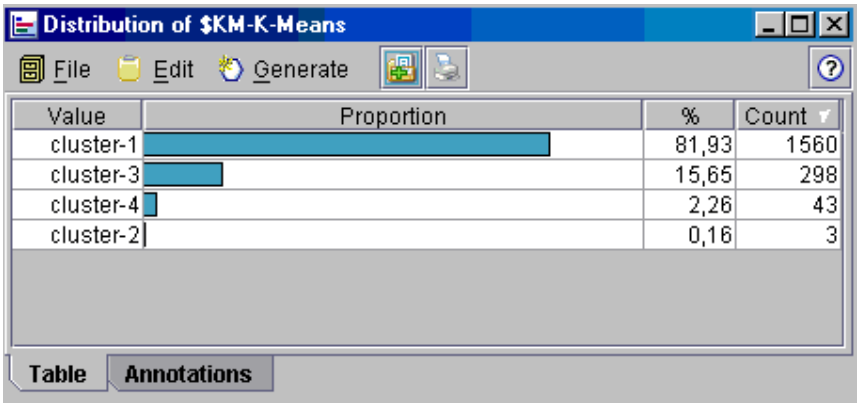

Figure 4

Concerning the 298 customers, characterized as «medium» by the K-Means Algorithm, relations between the payment they conduct through e-banking are sought. By the use of Apriori method and setting Minimum Rule Support = 10 and Minimum Rule Confidence = 20 five rules were determined and shown in Figure 5.

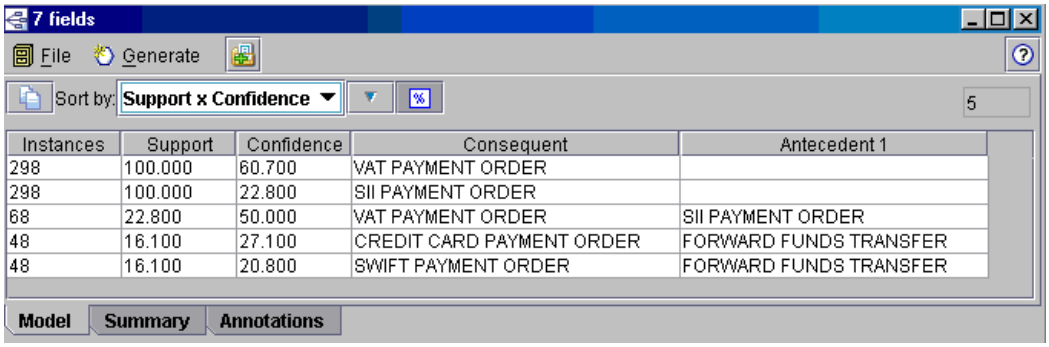

Figure 5

After observation between the rules obtained by the Apriori method it can be concluded that the most powerful is:

- *If* SII Payment Order *then* VAT Payment Order (confidence=50). (Rule 1)

Two other rules obtained exhibiting confidence > 20 are the following:

- *If* Forward Funds Transfers *then* Credit Card Payment Order (confidence=27,1), (Rule 2)
- *If* Forward Funds Transfers *then* Swift Payment Order (confidence=20,8), (Rule 3)

As seen, there exists strong relationship between VAT Payment Order and SII Payment Order. These strong relationships are visualized on the web graph of Figure 6.

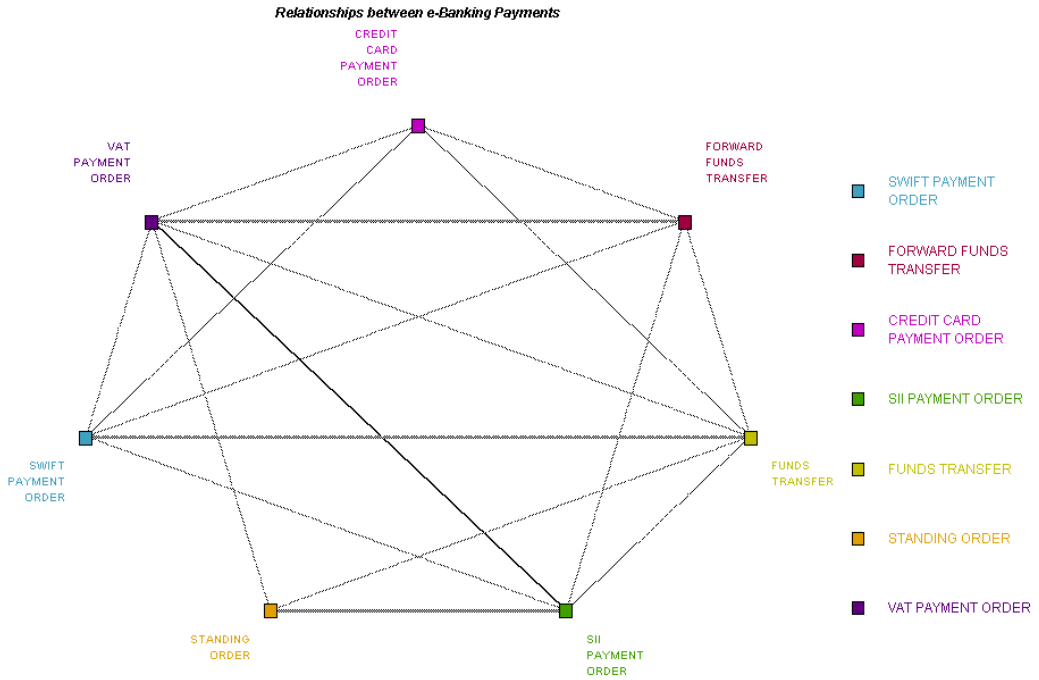

Figure 6

5. Conclusions and Future Work

In the present paper it is shown that the knowledge of RFM scoring of active e-banking users can rank them according to the pyramid model. The e-banking unit of a bank may easily identify the most important users-customers. The model continuously trained reveals also the way customers are transposed between different pyramid levels so that the bank administration has the opportunity to diminish customer leakage.

At the same time customer approach and new services and products promotion is improved since it is the bank's knowledge that it is more likely a customer to respond to a promotion campaign if this customer belongs to the 20% of more beneficial ones.

Correct recognition and analysis of the clustering results offers an advantage to the e-banking unit of a bank over the competition as it facilitates the procedure of decision support, especially bearing in mind that decisions should be based on justified analysis. Users-customers clustering could be subjected to further exploitation and research, such as discovery of association rules between different types of e-banking payment offered by a bank.

The basic outcome is that VAT Payment Orders and SII Payment Orders are the most popular and interconnected strongly. The detection of such relationships offers a bank a detailed analysis that can be used as a reference point for the conservation of the customer volume initially but also for approaching new customer groups (companies, freelancers) who conduct these payments in order to increase its customer number and profit.

Similar patterns can be derived after analysis of payment types of specific customer groups resulting from various criteria either gender, residence area, age, profession or any other criteria the bank assumes significant. The rules to be discovered show clearly the tendencies created in the various customer groups helping the bank to approach these groups as well as re-design the offered electronic services in order to become more competitive.

The use of other clustering algorithms as well as other data mining methods is a promising and challenging issue for future work. The application of RFM analysis can also be used in larger data sets, in order to produce completed results that will be updated continuously by training of the models.

In addition to the Apriori method used, detection of association rules can employ several other data mining methods such as decision trees in order to confirm the validity of the results.

6. References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1996) "Fast discovery of associations rules", *Advances in Knowledge Discovery and Data Mining*.

Bradley, P. and Fayyad, U. (1998) "Refining Initial Points for K-Means Clustering**",** *Proc. 15th International Conf. on Machine Learning*.

Brin S., Motwani, R., and Silverstein, C. (1997) "Beyond Market Baskets: Generalizing Association Rules to Correlations", *Proceedings ACM SIGMOD Conf. on Management of Data*.

Chen, M., Han, J. and Yu, P.(1997) "Data Mining: An Overview from Database Perspective", *Ieee Trans. On Knowledge And Data Engineering*.

Collier, K., Carey, B., Grusy, E., Marjaniemi, C., and Sautter, D. (1998) "A Perspective on Data Mining", *Northern Arizona University*

COMPAQ (2001) "Retain Customers and reduce risk", *White Paper.*

Curry, J. and Curry, A. (2000) "*The Customer Marketing Method: How to Implement and Profit from Customer Relationship Management.,*

DataPlus Millenium, (2001) "Data-Driven Analysis Tools and Techniques", *White Paper.*

Freitas, A. (1997) "A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction", *Genetic Programming 1997: Proceedings of the Second Annual Conference*

Freitas, A. (1999). "On Rule Interestingness Measures", *Knowledge-Based Systems journal*.

Freitas, A. (2002) "*A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery*", A. Ghosh and S. Tsutsui. (Eds.) Advances in Evolutionary Computation Springer-Verlag.

Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T.(1996) "Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization", *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

Han, E., Karypis, G., and Kumar, V. (1997) "Scalable Parallel Data Mining for Association Rules", *Proceedings of ACM SIGMOD Intl. Conf. on Management of Data.*

Hand, D, Mannila, H., Smyth P. (2001) "*Principles of Data Mining*". The MIT Press,

Hilderman, R. and Hamilton, H.(1999). "Knowledge Discovery and Interestingness Measures: A Survey", *Technical Report CS 99-04*, Department of Computer Science, University of Regina

Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2002) "*Data Mining of Associations Rules and the Process of Knowledge Discovery in Databases*".

Ikizler, N. and Guvenir, H.A. (2001) "Mining Interesting Rules in Bank Loans Data", *Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks*.

Im, K. and Park, S. (1999) "A Study on Analyzing Characteristics of Target Customers from Refined Sales Data", *APIEMS*.

Liu, J. and Kwok, J. (2000) "An Extended Genetic Rule Induction Algorithm", *Proceedings of the Congress on Evolutionary Computation (CEC).*

Madeira, S.A. (2002) "Comparison of Target Selection Methods in Direct Marketing", *MSc Thesis, Technical University of Lisbon*

Michail A. (2000) "Data Mining Library Reuse Patterns using Generalized Association Rules", *Proceedings of the 22nd International Conference on on Software Engineering*.

Ng, R., Lakshmanan, L. and Han. J. (1998) "Exploratory Mining and Pruning Optimizations of Constrained Association Rules", *Proceedings ACM SIGMOD International Conference on Management of Data*.

Smyth, P. and Goodman, R. (1991). "Rule Induction using Information Theory", *Knowledge Discovery in Databases 1991.*

SPSS (2002) "*Clementine 7.0 Users's Guide*". Integral solutions Limited.

SPSS (2001) "*Clementine Application Template for Customer Relationship Management 6.5*". Integral solutions Limited.

Srikant, R. and Agrawal, R. (1995). "Mining Generalized Association Rules", *Proceedings of the 21st Int'l Conference on Very Large Databases*.

Toivonen, H. (1996). "Discovery of Frequent Patterns in Large Data Collections", *Technical Report A-1996-5*, Department of Computer Science, University of Helsinki.

Yilmaz, T. and Guvenir, A. (2001). "Analysis and Presentation of Interesting Rules", *Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks*.

Zaki, M.J., Parthasarathy, S., Li, W., and Ogihara M. (1997) "Evaluation of Sampling for Data Mining of Association Rules", *7th Workshop Research Iss. Data Engg.*

Zha, H., Ding, C., Gu, M., He, X. and Simon, H. (2001) "Spectral Relaxation for K-means Clustering"*, Neural Info. Processing Systems.*