Clustering Analysis for Credit Default Probabilities in a Retail Bank Portfolio

Adela Ioana TUDOR, Adela BÂRA, Elena ANDREI (DRAGOMIR)
Bucharest Academy of Economic Studies
adela_sw@yahoo.com, bara.adela@ie.ase.ro, elena.andrei@gmail.com

Methods underlying cluster analysis are very useful in data analysis, especially when the processed volume of data is very large, so that it becomes impossible to extract essential information, unless specific instruments are used to summarize and structure the gross information. In this context, cluster analysis techniques are used particularly, for systematic information analysis. The aim of this article is to build an useful model for banking field, based on data mining techniques, by dividing the groups of borrowers into clusters, in order to obtain a profile of the customers (debtors and good payers). We assume that a class is appropriate if it contains members that have a high degree of similarity and the standard method for measuring the similarity within a group shows the lowest variance. After clustering, data mining techniques are implemented on the cluster with bad debtors, reaching a very high accuracy after implementation. The paper is structured as follows: Section 2 describes the model for data analysis based on a specific scoring model that we proposed. In section 3, we present a cluster analysis using K-means algorithm and the DM models are applied on a specific cluster. Section 4 shows the conclusions.
Keywords: *Data Mining, Cluster Analysis, Artificial Intelligence*

1 Introduction

Data mining is a technique that consists of analysing large volumes of information stored in data warehouses, in order to resolve decision problems. The technique is derived from three categories of software applications: the statistical ones, artificial intelligence applications based on neuro-fuzzy algorithms and the ones based on automated machine learning. Once the data has been prepared, the next step is to generate previously unknown patterns from the data using inductive learning. The most popular types of patterns are classification models, clustering and association rules that describe relations between attributes. Although methods and knowledge extraction techniques are applied in automatic mode, the process requires considerable human effort involved especially in the stages of analysis, but also in those of validation the results. There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework.

The steps taken in the entire process are [6]:

- collect data from multiple sources: web, text, databases, data warehouses ;
- data filtering by eliminating errors. When using a data warehouse, this process is removed because a process of extraction, transformation and loading (ETL) was already applied on the data ;
- establishing key data attributes that will participate in the DM process, by selecting those properties that interest the analysis ;
- application of templates and detection / analysis of new knowledge ;
- visualization, validation and evaluation of results.

The steps in the mining process are performed iteratively until meaningful business knowledge is extracted.

The main idea of the article is to build a model based on data mining techniques that can predict customer behaviour over time. We used hierarchical clustering method on

the set of records, in order to obtain a profile of the customers. Then, we applied the data mining models on the cluster with bad debtors, reaching a very high accuracy after implementation.

The paper is structured as follows: Section 2 describes the model for data analysis based on a specific scoring model that we proposed. In section 3, we present a cluster analysis using K-means algorithm and the DM models are applied on a specific cluster. Section 4 shows the conclusions.

For testing different methods we used Oracle Data Mining (ODM) that is organized around several generic operations, providing an unified interface for extraction and discovery functions. These operations include functions for construction, implementation, testing and manipulation of data to create models. ODM implements a series of algorithms for classification, prediction, regression, clustering, association, selection, and data analysis. Oracle Data Miner provides the following options for each stage: for transforming the data and build models (build), for testing the results (test) and for the evaluation and application on new data sets (apply).

2 The model for data analysis

The study is based on financial data in 2009 from an important bank in Romania and the target customers are credit card holders. Among the 18239 instances, 1489 record arrears. The research involves normalization of attribute values and a binary variable as response variable: 1 for the default situation and 0 for non-default. Explanatory variables or the attributes are found in the scoring model proposed by us that includes: Credit amount (the amount limit by which the debtor may have multiple withdrawals and repayments from the credit card), Credit balance, Opening date of the credit account, credit product identification, Category, Currency, Client's name, Gender, Age, Marital status, Profession, Client with history (including other banking products and payment history), Deposit state, Amounts of deposits opened in the bank, if applicable, Scoring rate (scoring points from 1 to 6, 1 is for the best and 6 for the weakest debtor).

Data was divided into two tables, one used for model construction and one for testing and validating the model. Each case contains a set of attributes, of which one is the profiling attribute, this attribute is called 'RESTANTIER' that means that the creditor is a bad payer if the value is 1 and is a good payer if the value is 0.

First we applied three data mining models on the set of instances: classification, Naïve Bayes and regression with support vector machines. We obtained a series of predictions for credits reimbursement and the comparative results show that only 845 of 18 239 records are incorrect predictions, representing 4.63% of total.

We also made a comparison of the incorrect predictions released by the three models: SVM registered 406 incorrect predictions (2,22%), NB recorded 703 incorrect predictions (3,85%), and LG registered only 324 wrong predictions (1,77%). From cost point of view, LG recorded the lowest cost, followed by SVM, and NB is detaching pretty much. Although, we can consider that all three models can be successfully applied in banking practice, we extended our study with a clustering analysis. The following section presents the results.

3 Cluster analysis

3.1. Clusters building

Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. In common parlance it is also called look-a-like groups. The simplest mechanism is to partition the samples using measurements that capture similarity or distance between samples [3].

Performing a cluster analysis targeting the classification of a set of objects, includes the following steps:

- choosing the characteristics subject to classification;
- choosing the measure in order to assess the proximity of objects;
- setting rules for grouping the classes or clusters;
- building the classes (ie classification of objects into classes);
- checking the consistency and significance of classification;
- choosing an optimal number of clusters, depending on the nature of the classification problem and the purpose;
- interpreting the significance of clusters.

We could say that cluster analysis can be understood as a classification technique or algorithm for organizing the data as classes or representative structures that verify certain properties. The results of cluster analysis are represented either by a single cluster solution, or cluster hierarchies, containing different ways of configuration of the objects in classes (ie cluster solutions).

We applied the clustering method on the set of records, taking into account only the important attributes. When building the model, we specified a number of 8 clusters, considering our objectives. Therefore, we took into consideration that the classified objects in each group to be as similar in terms of certain features and the classified objects into a group to differentiate as much as possible of the objects classified in any of the other groups.

The first criterion requires that each class to be as homogeneous compared to the characteristics considered for the classification of objects. The second criterion requires that each class may vary as much as possible in terms of classification features. A difficult problem that arises in cluster analysis is related to the need to assess the distances between classes or clusters.

K-means clustering is an iterative clustering method, and divides the data into a number of clusters by minimizing an error function which can be expressed. The *K-means* algorithm is a non-hierarchical approach to forming good clusters, used to group records based on similarity of values for a set of objects. Applying the concept allows classification for multiple classes and nonlinear relationships modelling between data (for prediction purposes). Even though, there may be difficulties often in establishing effective metrics, the technology being one of the few that accepts as input data of different nature (continuous, categorical, Boolean, etc.). Since the computing time is directly proportional to the number of instances in the data set, in the pre-processing stage, it is required to be selected from the original data set a subset of instances of reasonable size. This method proved to be effective in classification problems, when all associated key attribute classes have an equal representation as a percentage of the dataset. The algorithm based on k-NN technique allows only making an estimation of key attribute value, without generating additional information about the instances under review, the structure of the data set of classification categories of key attribute. This technique is used mainly in situations where for all attributes, the same function of distance is applied.

Figure 1 shows the cluster distribution and the number of observations in each cluster:

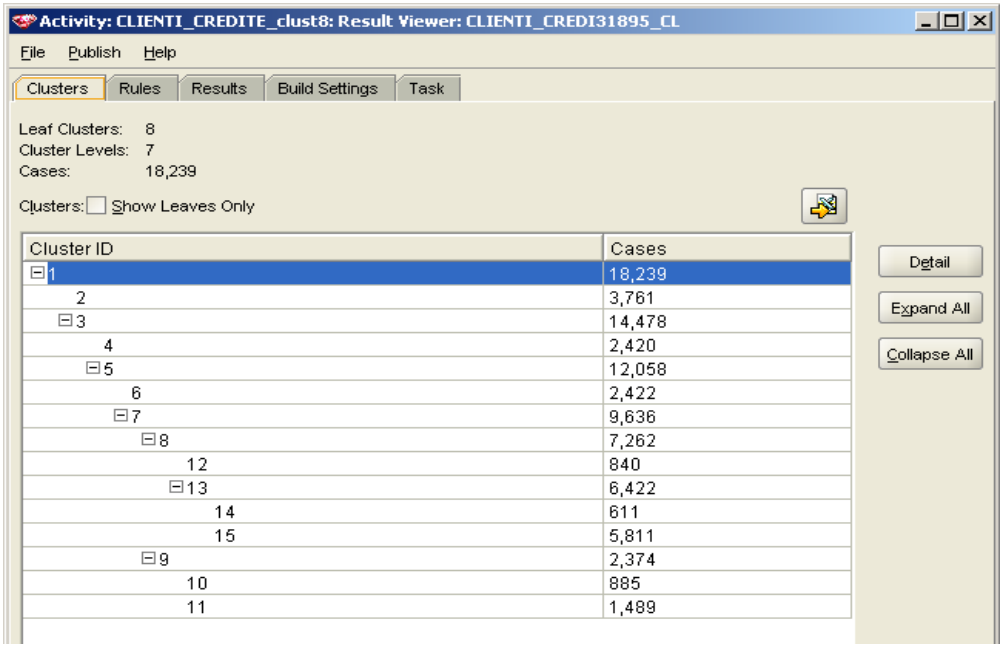

Fig. 1. Cluster distribution

Classification techniques allow us to specify by what combination of financial and demographic attributes can be characterized customers in each class and make predictions on the behaviour of new customers of the bank in order to include them in one class or another. In conclusion, in node 9 we found 2 leaves with 2 clusters, one with bad debtors and the other with good payers but with low scoring rate.

Frequency distributions of variables are represented by histograms. For example, in node 9, the amount of deposits is concentrated in the interval [600, 2142], and statistical events are focusing around 6, which means that most customers are characterized by the lowest score (ie those customers with a high probability of default). The histogram in Figure 2 shows the statistical distribution of events analyzed: 0 for the loan repayment and 1 for default. Thus, in node 9, good payers record 38% of all clients, while the debtors register the highest percentage (62%).

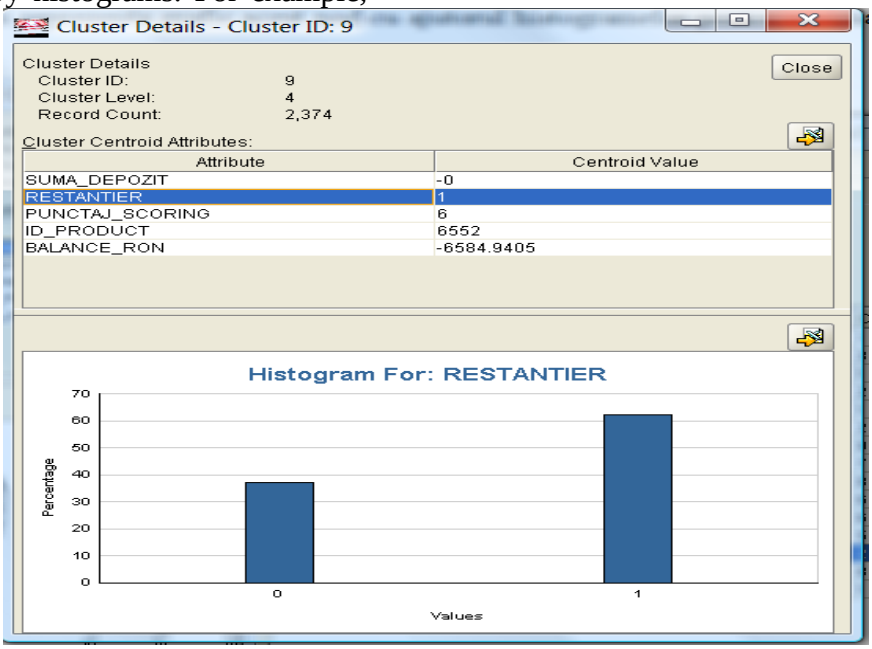

Fig. 2. The histogram for the debtor status

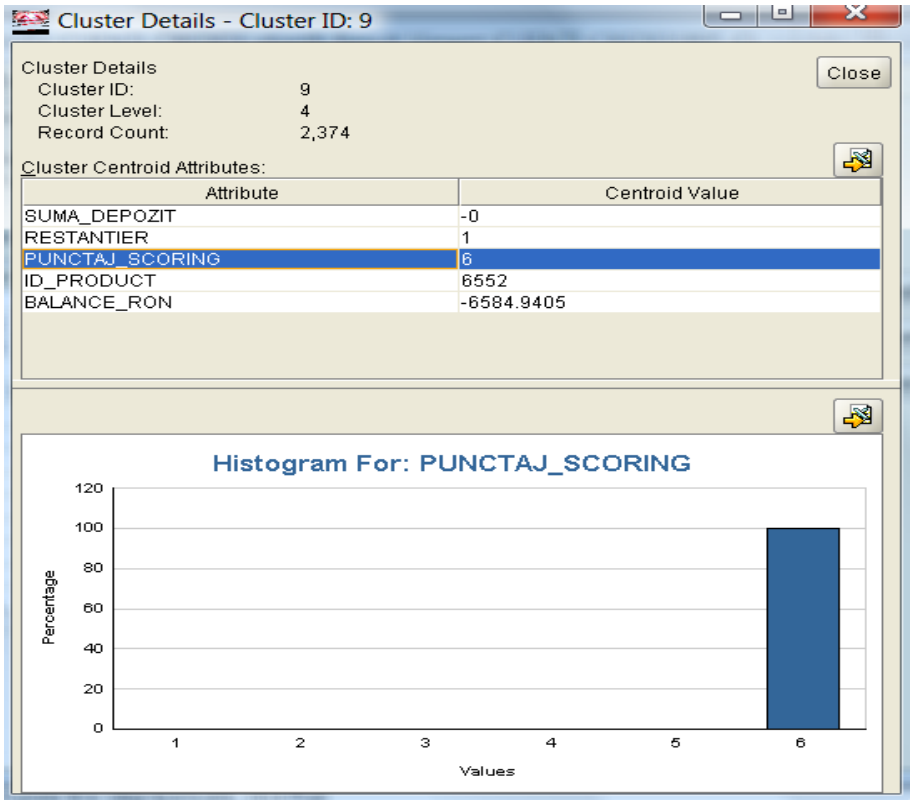

Fig. 3. The histogram for the scoring points

Figure 3 shows the distribution of the scoring rate in a node of customer data. Here, all the customers registered the weakest result when applying the scoring model. 100% received 6 points.

3.2. Applying the built model on clusters

Using cluster analysis, a customer 'type' can represent a homogeneous market segment. Targeting specific segments is an important issue of any bank's strategy. This way, each bank can develop and sell specific products and services oriented on clients needs and desires.

After applying the algorithms, we obtain the customers grouped in 8 clusters, as shown in the figure below:

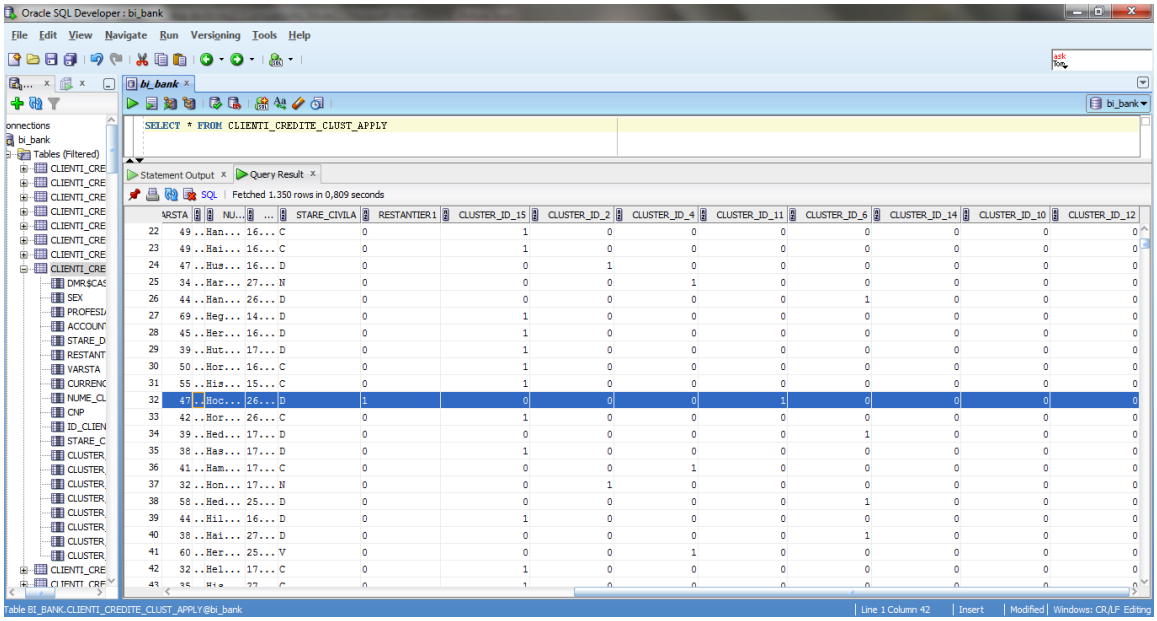

Fig. 4. Clusters situation

On these instances, we applied the predictive built models - Naive Bayes, SVM and LR, mentioning that we considered all the attributes and clusters obtained. The results are significant because all models registered over 99.8% accuracy.

Applying clustering method to our data, we identified the customers' profiles, depending on their credit history, account balance or scoring rate. Basically, the K-means algorithm was used to identify groups of customers based on scoring behaviour and divides the clients into three major classes: a) bad payers with no deposits or small amounts, who registered the lowest scoring rate (6 points) and big credit balance, b) potential insolvent clients, with scoring rate 5 or 6, small deposits and small credit balance and c) good payers with scoring rate 1-4 who have developed a good banking history.

4 Conclusions

We believe that, as of the management of massive amounts of data, data mining technology extracts successfully new knowledge from data collections, as shown in the survey presented. The built models provide decision alternatives for banking managers, based on modelling tools for the analysis of statistical data. Using statistical algorithms in order to determine patterns of behaviour, leads to creating some strong correlations for which the user is not able to generate queries. Classification and discovery of association rules are very important in the business decision process and management.

Analytical capabilities offered by the proposed solutions will produce relevant results and assist complex decision-making process, contributing significantly to improve performance in banking. Thus, an effective risk management is performed and also robust analysis, aiming to continuously improve achievement and profit margin.

Acknowledgements

This article is a result of the project POSDRU/88/1.5./S/55287 „Doctoral Programme in Economics at European Knowledge Standards (DOESEC)" . This project is co-funded by the European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies in partnership with West University of Timisoara. Also, this paper presents some results of the research project PN II, TE Program, Code 332: "Informatics Solutions for decision making support in the uncertain and unpredictable environments in order to

integrate them within a Grid network", financed within the framework of People research program.

References

[1] Han, J., Kamber, M. - *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001;

[2] Hattenschwiler, P. – Decision Support Systems, University of Fribourg, Department of Informatics, DS Group;

[3] Sambamoorthi N, *Hierarchical Cluster Analysis, Some Basics and Algorithms*, 2011.

[4] Ţiţan E, Tudor A, *Conceptual and statistical issues regarding the probability of default and modelling default risk* **,** Database Systems Journal, Vol. II, No. 1/2011, pg. 13-22.

[5] Tudor A., Bara A., Botha I. - *Solutions for analyzing CRM systems - data mining algorithms*, International Journal of Computers, Issue 4, Volume 5, 2011, pg. 485-493, ISSN: 1998-4308.

[6] Ullman, J. D. - *Data Mining Lecture Notes*, 2000.

Adela Ioana TUDOR has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2002. Two years later she graduated the MA in Financial Management and Capital Markets - DAFI, Faculty of Finance, Insurance, Banking and Stock Exchange Market. At present, she is a PhD candidate in the field of Economic Cybernetics and Statistics at the Academy of Economic Studies and also works for a leading multinational regional bank, having more than 10 years of professional experience in product management and financial analysis. Her major research interests are data mining optimization algorithms and solutions for financial institutions performance using statistical methods and techniques. During her research activity she published scientific papers and articles on OLAP technology and data mining models.

Adela BÂRA is a Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 7 books in the domain of economic informatics, over 40 published scientific papers and articles (among which over 20 articles are indexed in international databases, ISI proceedings, SCOPUS and 2 of them are ISI indexed). She participated (as director or as team member) in 4 research projects that have been financed from national research programs. She is a member of INFOREC professional association. From May 2009, she is the director of the Oracle Excellence Centre in the university, responsible for the implementation of the Oracle Academy Initiative program. Domains of competence: Database systems, Data warehouses, OLAP and Business Intelligence, Executive Information Systems, Decision Support Systems, Data Mining.

Elena DRAGOMIR has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2001, pursuing postgraduate studies in Quantitative Economics and Project Management. Currently, she is PhD candidate in the field of Economic Cybernetics and Statistics at the Academy of Economic Studies and also works for a major European leasing business, having more than 10 years of professional experience in credit risk assessment, underwriting

and management of risks and enterprise risk assessments. Her major research interests are credit risk minimization models and techniques for the leasing industry, estimation models of key risk parameters, stress-testing and enterprise risk management. During her research activity she published more than 10 scientific papers and articles on methods and techniques for risk assessment and management covering both the theoretical and practical aspects of the domain.