ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2007, Vol. 36, No. 1 A

# ON INTEGRATING UNSUPERVISED AND SUPERVISED CLASSIFICATION FOR CREDIT RISK EVALUATION

**Danuta Zakrzewska**

*Institute of Computer Science,*
*Technical University of Lodz, Poland*

**Abstract**. Credits granting are very important parts of banks' activities, as they may give big profits, but there is a big risk connected with making decisions in this area and mistakes may be very costly for financial institutions. The main idea in credit risk evaluation investigations consists of building classification rules that assign properly bank customers as good or bad payers. In the paper, the system based on combination of unsupervised and supervised classification is proposed. In the first step, by using clustering algorithm, clients are segmented into groups with similar features. In the second step, decision trees are built and classification rules, for each group of clients, are defined. To avoid redundancy, different attributes are taken into account during each kind of classification. The proposed approach allows for using different rules within the same data set, and for defining more accurately clients with high risk. The system was tested on the real credit-risk data sets. Some exemplary results concerning different groups of clients are presented.

## 1. Introduction

Decisions concerning credits granting are one of the most crucial in an every banks' policy. Well-allocated credits may become one of the biggest sources of profits for any financial organizations. On the other hand, this kind of bank's activity is connected with high risk as big amount of bad decisions may even cause bankruptcy. The key problem consists of distinguishing good (that surely repay) and bad (that likely default) credit applicants.

The main investigations, in this area, are based on building credit risk evaluation models, allowing for automating or at least supporting credit granting decisions. The research mainly focuses on adopting different classification techniques. Numerous methods, evaluating credit risk, were presented in the literature, so far. Most of them are based on traditional statistical methods like logistic regression [11], k-nearest neighbor [8], classification trees [5] or neural network models [6, 2, 13], as well as cluster analysis (see [4, 9, 10]). The performance of different classification algorithms as well as neural networks, together with accuracy of extracted models were broadly examined in [1] and [3].

Some of authors combined different models, to obtain strong general rules. In [12], authors built the decision system supporting evaluation of business credit applications, by applying integration of case based reasoning and decision rules. Such an approach allowed for connecting two kinds of representation knowledge and for formulating rules for a set of typical examples

In the paper, the combination of cluster analysis and decision tree models is investigated. This hybrid approach enables building rules for different groups of borrowers separately. In the first stage, bank customers are segmented into clusters, that are characterized by similar features and then, in the second step, for each group, decision trees are built to obtain rules that may indicate clients expected not to repay the loan. The main advantage of applying the integration of two techniques consists of building models that, may better predict risk connected with granting credits for each client, than while using each method separately.

The paper is organized as follows. First, there is presented the whole system architecture. Then data preparation process as well as each of applied techniques is described. In the third section, experiments on real credit data sets are presented and results obtained after each stage of the system are discussed. The final section presents concluding remarks.

## 2. The system architecture

The presented system, which aim is to support evaluation of credit risks, by building classification rules, is composed of three main steps. In the preprocessing phase, data preparation consists of identification of attributes to use during the next steps.

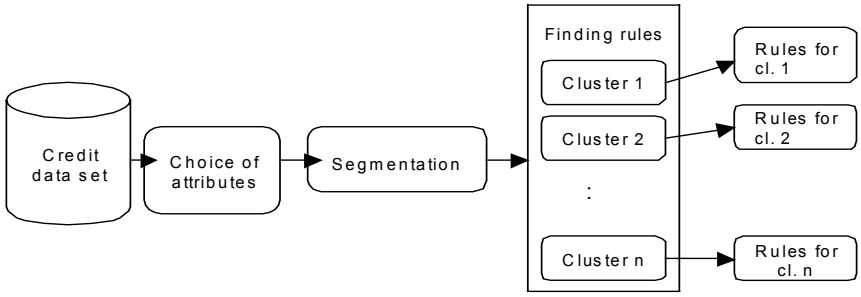


**Figure 1**. System architecture

The attributes are divided into two separate groups. The first one is applied in the next step to segment clients with similar features. The second group of attributes, in turn, is used to build classification rules, for each cluster of customers, in the final stage.

Each new applicant is assigned to one of the clusters and the decision concerning credit's granting is taken in accordance with rules generated for it. The overview of the system is presented in Figure 1.

### 2.1. Data preparation

During this stage credit data attributes are divided into two groups. The first one is used in cluster analysis for segmenting data, the second one will be applied later, while building classification rules for each cluster.

Financial institutions use different attributes in collected credit data. Generally they may have quantitative or qualitative character. Examples of both kinds of attributes are presented in Table1 and Table2.

**Table 1**. Examples of quantitative attributes

| No | Attribute name |
|---|---|
| 1 | Term |
| 2 | Credit amount |
| 3 | Age |
| 4 | Deposit amount |
| 5 | Payment rate |
| 6 | Number of years employed |
| 7 | Income |

Two aspects are important at this step. Attributes should fit to classification techniques. In case of decisions trees all quantitative (continuous) values should be changed into qualitative (nominal). On the other hand, one should be very careful, while choosing nominal attributes for using clustering techniques, as only special distance functions may work properly for variables of this type (see [7]).

At this stage of the system a decision of expert, which attributes should be chosen, for every step, is necessary.

**Table 2.** Examples of qualitative attributes

| No | Attribute name |
|---|---|
| 1 | Checking account |
| 2 | Credit history |
| 3 | Purpose |
| 4 | Savings account |
| 5 | Present employment |
| 6 | Installment rate |
| 7 | Personal status and sex |
| 8 | Other parties |
| 9 | Present residence since |
| 10 | Property |
| 11 | Other installments |
| 12 | Housing |
| 13 | Number of existing credits at this bank |
| 14 | Job |
| 15 | Number of dependents |
| 16 | Telephone |
| 17 | Foreign worker |

### 2.2. Segmenting customers

Cluster analysis techniques become very popular in customer segmentation area. In banking, customer segmentation allows not only reducing exposure to credit risk, but also matching campaigns to customers and personalizing services according to client interests. In the paper, the focus is based on the first purpose, however, one can also achieve the others mentioned above, by using only one of the stages of the system.

One of the main advantage of the clustering technique is that it does not assume any specific distribution on the data, so it is suitable for credit risk analysis [10]. The main disadvantage of the method consists on big dependence of experts' opinions in many cases.

Cluster analysis techniques have been broadly investigated in the literature (see [7] for example). The comparisons of performance of different algorithms for bank customer segmentation have been discussed with details in [15]. For the presented system, well

known, k-means algorithm has been chosen, because of its simplicity and efficacy on big data sets. However the method depends significantly on the initial assignments, what may entail in not finding the most optimal cluster allocation at the end of the process, but as it was concluded in [15], k-means is very efficient for large multidimensional data sets. Besides, tests at the early stage of building the system showed its supremacy on agglomerative hierarchical clustering algorithms that did not give satisfying results, especially in the case of noise presence.

The segmentation module is adjusted into clustering by numerical attributes and as these may have different range of values, it is enhanced in normalization procedure. The distance between objects (customers) can be calculated by the most common Manhattan or Euclidean metrics.

### 2.3. Building decision rules

In this step, well known, C4.5 algorithm is used. It is based on ID3 decision tree induction algorithm enhanced with improvements concerning dealing with numeric attributes, missing values, noisy data, and generating rules from trees (see [14]). This technique is also equipped with tree pruning mechanism.

Classification and decision rules induction are done for every cluster found in the previous stage of the system. Credit risk is evaluated for different groups of borrowers separately, as each rule is generated only on data of customers assigned to one cluster. Experts may even use different choice of attributes for different segments of clients.

Assessment of classification accuracy is done by calculating the percentage of correctly classified instances and by estimating complexity of generated decision trees. The last one is expressed by the number of leaf nodes and the size of obtained tree expressed by total number of nodes. Especially this feature of the decision rule is very important as experts look for clear and simple rules. If the ratio of correctly classified instances is comparable, the complexity should be the main factor deciding on the chosen rule.

## 3. Experiments

Experiments were done on the real life credit risk data sets: German bank data available at *http://www.stst.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html*, and Japan bank data, that can be found at *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit*, each of which with different attributes. Main experiments consist of evaluating and comparing the quality of results obtained by building decision rules on different segments of users separately with those received while using the whole credit risk data set.

### 3.1. Case one

German bank credit data set contains records, of clients who granted or failed in credit applications described by 21 attributes. Three of them: term, credit amount and age are numeric, while 17 the others are qualitative (all of them are presented in Table 2). Additional attribute *class* is nominal (0,1) and means the decision if credit is granted.

During experiments, all quantitative data were used in the cluster analysis. After several tests clients were segmented into four groups, that may be characterized as follows: the first one of rather young people with big credit amount and long term of repayment, the second one of middle-aged persons with average credit amount and average term of repayment; third group of young people, with low credit amount and rather short term of repayment and the last one of old persons with average credit amount and average term of repayment. Cluster centers for all the groups are presented in Table 3.

**Table 3.** Cluster centers attributes values

| Cluster | Age | Credit amount (in DM) | Credit repayment (in months} |
|---|---|---|---|
| Cluster1 | 32 | 4773 | 40 |
| Cluster2 | 36 | 3197 | 20 |
| Cluster3 | 30 | 1733 | 13 |
| Cluster4 | 57 | 3653 | 22 |

Decision rules were built, in two ways, by starting with the set of all 17 attributes and by using different attributes for each cluster. The best results were obtained in the second case, what can be easily seen in Table 5 and Table 6. However, the complexity of obtained decision trees are the same, but the rules are formulated by using different attributes, and the numbers of correctly classified instances, in the second case, are significantly greater than in the first one.

The rules received for each cluster separately are significantly less complex than the ones obtained for all data. Table 4 presents the Decision Table visualizing the decision rules for all the data, we can see that three attributes: *checking account, savings account* and *foreign worker* are used in that tree, while for each cluster only one attribute is necessary to build the rule. For example, if we consider the group of young people with big credit amount and long term of repayment the *checking account* attribute occurred to be crucial while for second and third clusters *other installment* value were deciding.

### 3.2. Case two

Now, there will be considered Japan bank credit data set, that also contains records, of clients who granted or failed in credit applications. Data records are described by 11 attributes (see Table 7) including *class.* All of them have more demographic character

than German data. What is more, number of five numeric attributes, which makes half of all of them, allows for distributing equally the weight of the decision process between both stages of the system.

**Table 4**. Decision Table for the rules extracted for all data (amounts in DM)

| Checking account | ≤ 0 | 0≤ ...<200 | ≥200 | No account | | | | |
|---|---|---|---|---|---|---|---|---|
| Savings account (all assets) | - | - | - | <100 | 100≤ ...<500 | 500≤..<1000 | ≥1000 | No savings |
| Foreign worker | - | - | - | - | - | - | - | Y | N |
| Class | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |

**Table 5.** Comparison of classification accuracy rules built starting with all the attributes

| Data Set | Number of leaves | Size of the tree | Correctly classified instances |
|---|---|---|---|
| Cluster1 | 4 | 5 | 71% |
| Cluster2 | 3 | 4 | 56% |
| Cluster3 | 3 | 4 | 54% |
| Cluster4 | 2 | 3 | 47% |
| All data | 9 | 12 | 61% |

**Table 6.** Comparison of classification accuracy rules built on different attributes

| Data Set | Number of leaves | Size of the tree | Correctly classified instances |
|---|---|---|---|
| Cluster1 | 4 | 5 | 86% |
| Cluster2 | 3 | 4 | 65% |
| Cluster3 | 3 | 4 | 70% |
| Cluster4 | 2 | 3 | 79% |
| All data | 9 | 12 | 70% |

**Table 7**. Japanese bank data attributes

| No | Attribute name | Attribute type |
|---|---|---|
| 1 | Class | nominal |
| 2 | Unemployed | nominal |
| 3 | Purpose | nominal |
| 4 | Sex | nominal |
| 5 | Single/married | nominal |
| 6 | Problematic region | nominal |
| 7 | Age | numeric |
| 8 | Account balance | numeric |
| 9 | Payment rate | numeric |
| 10 | Credit repayment in months | numeric |
| 11 | Number of years employed | numeric |

Also in this case, in the first step, the number of four clusters was chosen to divide clients into groups according to all numeric attributes. It is worth to notice that *credit rate* instead of *credit amount* is registered as an attribute. Customers assigned into the first cluster are rather young, with average account balance, employed for rather long time with long term credit repayment, while those assigned into the third cluster are characterized by short term of working and short term of credit repayment. The second and the forth groups contain data of rather old clients. Those assigned into the second cluster are well situated, employed for a long time with long term of credit repayment, while those assigned into the cluster number four have rather low account balance but also low payment rate and short repayment term (see Table 8).

**Table 8.** Cluster centers attributes values

| Cluster | Age | Account balance | Payment rate | Credit repayment ( months) | Number of years employed |
|---|---|---|---|---|---|
| Cl.1 | 33 | 83 | 11 | 20 | 14 |
| Cl.2 | 49 | 131 | 49 | 23 | 23 |
| Cl.3 | 27 | 61 | 9 | 10 | 4 |
| Cl.4 | 49 | 46 | 7 | 9 | 10 |

All the rules built in the second stage are very simple (see Table 9). Decision trees constructed for clusters number one and four are the same as for the set of all data and depend only on the one attribute *unemployed*. All the instances contained in the cluster number two should be classified as yes, with the highest precision (89%). For the cluster number three the system indicated the simple tree with the attribute *problematic region*. However the accuracy for the rules determined for this cluster are less than for rules built for the whole data set, but those based on the attribute *unemployed*, give even less, for this cluster: 71% of correctly classified instances.

**Table 9.** Comparison of classification accuracy

| Data Set | Number of leaves | Size of the tree | Correctly classified instances |
|---|---|---|---|
| Cluster1 | 2 | 3 | 68% |
| Cluster2 | 1 | 1 | 89% |
| Cluster3 | 2 | 3 | 73% |
| Cluster4 | 2 | 3 | 86% |
| All data | 2 | 3 | 76% |

### 3.3. Remarks

Data sets chosen for experiments, were rather small (about 100 – 125 instances each), to ensure full control on the whole process. But the investigations were also done for much bigger data sets, which count 1000 instances and more.

During first stage, results for different number of required clusters, together with choice of distance functions were examined. After some trial computations the number of four clusters was chosen as optimal, however it may be different for various data sets. In all considered cases Manhattan and Euclidean functions gave similar results.

In the second stage for all the considered cases models were built on the full training set by using as the test mode: 10 fold cross validation. C4.5 technique gave much better results, measured by accuracy and simplicity of constructed rules, than Id3 decision tree technique.

## 4. Conclusions

In the paper a possibility of connecting unsupervised and supervised techniques for credit risk evaluation is investigated. The presented technique allows for building different rules for different groups of customers. In the proposed approach, each credit applicant is assigned to the most similar group of clients from the training data set and credit risk is evaluated by applying the rules proper for this group.

Results obtained on the real credit risk data sets showed higher precisions and simplicity of rules obtained for each cluster than for rules connected with the whole data set.

Future research will focus on further investigations of both stages of the system, especially by improving clustering method, including possibility of segmenting according to attributes of nominal or mixed types.

## References


**[1]** **B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen**. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, 54, 2003, 627-635.

**[2]** **B. Baesens, R. Setieno, Ch. Mues, J. Vanthienen**. Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science,* 49(3), 2003, 312-329.

**[3]** **M. Bensic, N. Sarlija, M. Zekic-Susac.** Modelling Small-Business Credit Scoring by Using Logistic Regression. Neural Networks and Decision Trees. *Intelligent Systems in Accounting, Finance and Management,* 13, 2005, 133-150.

**[4]** **G. Chi, J. Hao, Ch. Xiu, Z. Zhu.** Cluster Analysis for Weight of Credit Risk Evaluation Index. *Systems Engineering-Theory Methodology, Applications,* 10(1), 2001, 64-67.

**[5]** **R.H. Davis, D.B. Edelman, A.J. Gammerman.** Machine learning algorithms for credit-card application. *IMA Journal of Management Mathematics,* 4, 1992, 43-51.

**[6]** **V.S. Desai, J.N. Crook, G.A. Overstreet Jr.** On comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* 95(1), 1996, 24-37.

**[7]** **J. Han, M. Kamber**. Data Mining: Concepts and Techniques. *Morgan Kaufmann Publisher*s, 2001.

**[8]** **W.E. Henley, D.E. Hand.** Construction of a k-nearest neighbor credit-scoring system. *IMA Journal of Management Mathematics,* 8, 1997, 305-321.

**[9]** **M. Lundy**. Cluster Analysis in Credit Scoring. Credit Scoring and Credit Control. *New York: Oxford University Press,* 1993.

**[10]** **Y.-Z. Luo, S.-L. Pang, S.-S. Qiu**. Fuzzy Cluster in Credit Scoring. *Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an,* 2-5 November 2003, 2731-2736.

**[11]** **A. Steenackers, M.J. Goovaerts.** A credit scoring model for personal loans. *Insurance Mathematics & Economics,* 8, 1989, 31-34.

**[12]** **J. Stefanowski, S. Wilk.** Evaluating Business Credit Risk by Means of Approach – Integrating Decision Rules and Case-Based Learning. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 10, 2001, 97-114.

**[13]** **D. West.** Neural network credit scoring models. *Computers & Operations Research*, 27, 2000, 1131-1152.

**[14]** **I.H. Witten, E. Frank,** Practical Machine Learning Tools and Techniques with Java Implementations. *Morgan Kaufmann Publishers,* 1999.

**[15]** **D. Zakrzewska, J. Murlewski.** Clustering Algorithms for Bank Customer Segmentation. *Proceedings of 5th International Conference on Intelligent Systems Design and Applications ISDA*'05, *IEEE Computer Society*, 8-10 *September* 2005, *Wroclaw Poland*, 197-202.


Received March 2007.