BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXII No. 3/2010	66 - 75	Seria Ştiinţe Economice

Applying TwoStep Cluster Analysis for Identifying Bank Customers' Profile

Daniela Şchiopu

Petroleum-Gas University of Ploiesti, Informatics Department, 39 Bucuresti Blvd., Ploieşti, Romania
e-mail: daniela_schiopu@yahoo.com

Abstract

In this paper we analyze information about the customers of a bank, dividing them into three clusters, using SPSS TwoStep Cluster method. This method is perfect for our case study, because, compared to other classical clustering methods, TwoStep uses mixture data (both continuous and categorical variables) and it also finds the optimal number of clusters. TwoStep creates three customers' profiles. The largest group contains skilled customers, whose purpose of the loan is education or business. The second group consists in persons with real estate, but mostly unemployed, which asked for a credit for retraining or for household goods. The third profile gathers people with unknown properties, who make a request for a car or a television and then for education. The benefit of the study is reinforcing the company's profits by managing its clients more effectively.

Key words: *TwoStep Cluster, clustering, pre-clustering, CF tree*

JEL Classification: *C63, C46, C19*

Introduction

The applications that can use clustering algorithms belong to various fields. However, most of these algorithms work with numerical data or categorical data. Nevertheless, data from real world contains both numerical and categorical attributes. TwoStep Cluster is an SPSS method which solves this problem.

In the present paper, we intend to identify the bank customers' profiles, starting with a public dataset provided by a German bank and using TwoStep Cluster. This method has the advantage of determining the proper number of clusters, so the aim is to find this number of profiles, for managing the existing and the possible clients effectively.

In the following sections, we introduce the TwoStep Cluster method and our case study with inputs, outputs and the interpretation of the results.

Statistical Approach

Data grouping (or data clustering) is a method that can form classes of objects with similar characteristics. Clustering is often confused with classification, but there is a major difference between them, namely, when classifying, the objects are assigned to predefined classes, whereas in the case of clustering, those classes must be defined too.

Clustering techniques are used when we expect the data to group together naturally in various categories. The clusters are categories of items with many features in common, for instance, customers, events etc. If the problem is complex, before clustering the data, other data mining techniques can be applied (such as neural networks or decision trees).

Classical methods of clustering use hierarchical or partitioning algorithms. The hierarchical algorithms form the clusters successively, on the basis of clusters established before, while the partitioning algorithms determine all the clusters at the same time, building different partitions and then evaluating them in relation to certain criteria.

In SPSS[1], clustering analysis can be performed using TwoStep Cluster, Hierarchical Cluster or K-Means Cluster, each of them relying on different algorithm to create the clusters. The last two are classical methods of classification, based on hierarchical, respectively partitioning algorithms, while TwoStep method is especially designed and implemented in SPSS.

In terms of types of data considered for application, Hierarchical Cluster is limited to small datasets, K-Means is restricted to continuous values and TwoStep can create cluster models based on both continuous and categorical variables.

Next, we approach the TwoStep method, highlighting its advantages in the field under discussion.

The TwoStep Cluster Analysis

TwoStep Cluster is an algorithm primarily designed to analyze large datasets. The algorithm groups the observations in clusters, using the approach criterion[2]. The procedure uses an agglomerative hierarchical clustering method[3]. Compared to classical methods of cluster analysis, TwoStep enables both continuous and categorical attributes. Moreover, the method can automatically determine the optimal number of clusters.

TwoStep Cluster involves performing the following steps:

- pre-clustering;
- solving atypical values (outliers) - optional;
- clustering.

In the *pre-clustering step*, it scans the data record one by one and decides whether the current record can be added to one of the previously formed clusters or it starts a new cluster, based on the distance criterion[4]. The method uses two types of distance measuring: Euclidian distance and log-likelihood distance[5].

Pre-clustering procedure is implemented by building a data structure called CF (cluster feature) tree, which contains the cluster centers. The CF tree consists of levels of nodes, each node having a number of entries. A leaf entry is a final sub-cluster. For each record, starting from the root node, the nearest child node is found recursively, descending along the CF tree. Once reaching a leaf node, the algorithm finds the nearest leaf entry in the leaf node. If the record is within a threshold distance of the nearest leaf entry, then the record is added into the leaf entry and the CF tree is updated. Otherwise, it creates a new value for the leaf node. If there is enough

[1] *** SPSS (Statistical Package for the Social Sciences), available at http://www.spss.com/ , [accessed on 3 July 2010].
[2] *** *Analiza datelor*, available at http://www.spss.ro/detail.php?id=18 , [accessed on 20 July 2010].
[3] *** *The SPSS TwoStep cluster component*, Technical report, available at http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf , [accessed on 20 July 2010].
[4] ibidem
[5] Arminger, G., Clogg, C., Sobel, M., *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York, 1995, pp. 130.

space in the leaf node to add another value, that leaf is divided into two values and these values are distributed to one of the two leaves, using the farthest pair as seeds and redistributing the remaining values based on the closeness criterion.

In the process of building the CF tree, the algorithm has implemented an optional step that allows solving atypical values (outliers). Outliers are considered records that do not fit well into any cluster. In SPSS, the records in a leaf are considered outliers if the number of records is less than a certain percentage of the size of the largest leaf entry in the CF tree; by default, this percentage is 25%. Before rebuilding the CF tree, the procedure searches for potential atypical values and puts them aside. After the CF tree is rebuilt, the procedure checks if these values can fit in the tree without increasing the tree size. Finally, the values that do not fit anywhere are considered outliers.

If the CF tree exceeds the allowed maximum size, it is rebuilt based on the existing CF tree, by increasing the threshold distance. The new CF tree is smaller and allows new input records.

The *clustering stage* has sub-clusters resulting from the pre-cluster step as input (without the noises, if the optional step was used) and groups them into the desired number of clusters. Because the number of sub-clusters is much smaller than the number of initial records, classical clustering methods can be used successfully. TwoStep uses an agglomerative hierarchical method which determines the number of clusters automatically.

Hierarchical clustering method refers to the process by which the clusters are repeatedly merged, until a single cluster groups all the records. The process starts with defining an initial cluster for each sub-cluster. Then, all clusters are compared and the pair of clusters with the smallest distance between them is merged into one cluster. The process repeats with a new set of clusters until all clusters have been merged. Thus, it is quite simple to compare the solutions with a different number of clusters.

To calculate the distance between clusters, both the Euclidian distance and the log-likelihood distance can be used.

The Euclidian distance can be used only if all variables are continuous. The Euclidian distance between two points is defined as the square root of the sum of the squares of the differences between coordinates of the points[6]. For clusters, the distance between two clusters is defined as the Euclidian distance between their centers. A cluster center is defined as the vector of cluster means of each variable[7].

Log-likelihood distance can be used both for continuous and categorical variables. The distance between two clusters is correlated with the decrease of the natural logarithm of likelihood function, as they are grouped into one cluster. To calculate log-likelihood distance, it is assumed that the continuous variables have normal distributions and the categorical variables have multinomial distributions, and also the variables are independent of each other.

The distance between clusters *i* and *j* is defined as[8]:

$$d(i, j) = \xi_i + \xi_j - \xi_{<i,j>} \tag{1}$$

where:

[6] *** *Euclidian distance*, available at http://en.wiktionary.org/wiki/Euclidean_distance [accessed on 22 July 2010].
[7] *** *TwoStep Cluster Analysis*, available at http://support.spss.com/productsext/spss/documentation/statistics/algorithms/14.0/twostep_cluster.pdf [accessed on 22 July 2010].
[8] ibidem

$$\xi_s = -N_s\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^2+\hat{\sigma}_{sk}^2\right)+\sum_{k=1}^{K^B}\hat{E}_{sk}\right) \tag{2}$$

and in equation (2),

$$\hat{E}_{sk} = -\sum_{l=1}^{L_k}\frac{N_{skl}}{N_s}\log\frac{N_{skl}}{N_s} \tag{3}$$

with notations:

$d(i,j)$ is the distance between clusters *i* and *j*; $<i,j>$ index that represents the cluster formed by combining clusters *i* and *j*; K^A is the total number of continuous variables; K^B is total number of categorical variables; L_k is the number of categories for the *k*-th categorical variable; N_s is the total number of data records in cluster *s*; N_{skl} is the number of records in cluster *s* whose categorical variable *k* takes *l* category; N_{kl} is the number of records in categorical variable *k* that take the *l* category; $\hat{\sigma}_k^2$ - the estimated variance (dispersion) of the continuous variable *k*, for the entire dataset; $\hat{\sigma}_{sk}^2$ - the estimated variance of the continuous variable *k*, in cluster *j*.

To determine the number of clusters automatically, the method uses two stages. In the first one, the indicator BIC (Schwarz's Bayesian Information Criterion) or AIC (Akaike's Information Criterion) is calculated for each number of clusters from a specified range; then this indicator is used to find an initial estimation for the number of clusters.

For *J* clusters, the two indicators are computed according to equations (4) and (5), as follows[9]:

$$BIC(J) = -2\sum_{j=1}^{J}\xi_j + m_J\log(N), \tag{4}$$

$$AIC(J) = -2\sum_{j=1}^{J}\xi_j + 2m_J, \tag{5}$$

where:

$$m_J = J\left\{2K^A+\sum_{k=1}^{K^B}(L_k-1)\right\} \tag{6}$$

The relative contribution of variables to form the clusters is computed for both types of variables (continuous and categorical).

For the continuous variables, the importance measure is based on:

$$t = \frac{\hat{\mu}_k-\hat{\mu}_{sk}}{\hat{\sigma}_{sk}}\sqrt{N_k} \tag{7}$$

where:

$\hat{\mu}_k$ is the estimator of *k* continuous variable mean, for entire dataset, and $\hat{\mu}_{sk}$ is the estimator of *k* continuous variable mean, for cluster *j*.

In H_0 (the null hypothesis), the importance measure has a Student distribution with $N_k - 1$ degrees of freedom. The significance level is two-tailed.

[9] ibidem

For the categorical variables, the importance measure is based on χ^2 test:

$$\chi^2 = \sum_{l=1}^{L_k} \left(\frac{N_{skl}}{N_{kl}} - 1 \right)^2, \tag{8}$$

which, in null hypothesis, is distributed as a χ^2 with L_k degrees of freedom.

Regarding the cluster membership of the items, the records are allocated on the specifications of resolving atypical values (the noises) and the options for measuring the distances.

If the option of solving the atypical values is not used, the values are assigned to the nearest cluster, according to the method of distance measuring. Otherwise, the values are treated differently, as follows:

- in the case of the Euclidian method, an item is assigned to the nearest cluster if the distance between them is smaller than a critical value,

$$C = 2\sqrt{\frac{1}{JK^A} \sum_{j=1}^{J} \sum_{k=1}^{K^A} \hat{\sigma}_{jk}^2} \tag{9}$$

  Otherwise, the item is declared as noise (outlier).

- If the log-likelihood method is chosen, it assumes the noises follow a uniform distribution and it computes both the log-likelihood corresponding to assigning an item to a noise cluster and that resulting from assigning it to the nearest non-noise cluster. Then, the item is assigned to the cluster that has obtained the highest value of logarithm. This is equivalent to assigning an item to the nearest cluster if the distance between them is smaller than a critical value. Otherwise, the item is designated as noise.

In conclusion, an important advantage of the method is that it operates with mixed data (both continuous and categorical data). Another advantage is that, although the TwoStep method works with large datasets, in terms of time required for processing such data, this method needs a shorter time than other methods[10]. As a disadvantage, the TwoStep method does not allow missing values and the items that have missing values are not considered for analysis.

Case Study

Since TwoStep Cluster is often preferred first for large datasets and second for handling mixture data, we applied this method using some public data referring to the clients of a bank for clustering this data. (On the other hand, some of this data was used in another application to reduce the dimensionality applying PCA – Principal Component Analysis)[11]. The input and the output of this method are presented further below.

A related paper presents a study for control client using the same method which we used in this paper[12]. The authors propose a policy for consolidating a company's profits by selecting the clients using the cluster analysis method of CRM (Client Relationship Management), managing the resources better. For the realization of a new service policy, they analyze the level of contribution of the clients' service pattern: total number of visits to the homepage, service type,

[10] Bacher, J., Wenzig, K., Vogler, M., *SPSS TwoStep Cluster – A First Evaluation*, available at http://www.statisticalinnovations.com/products/twostep.pdf , [accessed 23 July 2010].
[11] Ioniță, I., Şchiopu, D., Using principal component analysis in loan granting, *Bulletin of Petroleum-Gas University of Ploieşti*, Mathematics, Informatics, Physics Series, vol. LXI, no. 1/2010.
[12] Park, H.-S, Baik, D.-K., A study for control of client value using cluster analysis, *Journal of Network and Computer Applications*, Vol. 29, No. 4, Elsevier, 2006, pp. 262-276.

service usage period, total payment, average service period, service charge per homepage visit and profits through the cluster analysis of clients' data. The clients were grouped into four clusters according to the contribution level in terms of profits.

Input

The dataset that has been used for our case study has been obtained from a public database that contains credit data of a German bank[13]. The dataset has 1000 records and is presented in a table in SPSS. This table contains information about the duration of the credit, credit history, purpose of the loan, credit amount, savings account, years employed, payment rate, personal status, residency, property, age, housing, number of credits at bank, job, dependents and credit approval. In Table 1 we present part of this data.

Table 1. Source data

Duration	CreditHistory	Purpose	CreditAmount	YearsEmployed	PaymentRate	PersonalStatus
6	critical	television	1169.0	>=7	4.0	male_single
48	ok_til_now	television	5951.0	<4	2.0	female
12	critical	education	2096.0	<7	2.0	male_single
42	ok_til_now	furniture	7882.0	<7	2.0	male_single
24	past_delays	car_new	4870.0	<4	3.0	male_single
36	ok_til_now	education	9055.0	<4	2.0	male_single
24	ok_til_now	furniture	2835.0	>=7	3.0	male_single
36	ok_til_now	car_used	6948.0	<4	2.0	male_single
12	ok_til_now	television	3059.0	<7	2.0	male_divorced
30	critical	car_new	5234.0	unemployed	4.0	male_married

The database contains 9 categorical variables and 7 continuous variables. Continuous variables are standardized by default. Because we use mixture data, we have only log-likelihood option for distance measure.

In the first running, we choose BIC to determine the number of clusters, though we may override this and specify a fixed number. The results obtained using AIC running are not different from those obtained with BIC, so below we present only those obtained with BIC indicator.

Regarding the noises (the outliers) from our dataset, we do not check the noise handling option. Outliers are defined as cases in CF tree, in other leaves with fewer than the specified percentage of the maximum leaf size.

An important option given by SPSS is to export in XML format the CF tree or the entire model. This allows the model to be updated later for additional datasets.

Output

The *Auto-Clustering* statistics table in SPSS output can be used to assess the optimal number of clusters in our analysis, as shown in Table 2.

[13] *** *Source data*, available at http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ [accessed on 9 May 2010].

Table 2. Auto-Clustering

Number of clusters	Schwarz's Bayesian Criterion (BIC)	Ratio of Distance Measures
1	26154,864	
2	25113,605	1,438
3	24488,152	1,542
4	24196,817	1,210
5	24012,292	1,189
6	23908,626	1,185
7	23871,920	1,131
8	23877,138	1,062
9	23900,958	1,009
10	23927,369	
11	23982,083	1,125
12	24066,712	1,047
13	24162,160	1,002
14	24258,157	1,030
15	24360,773	1,003

In Table 2, although the lowest BIC coefficient is for seven clusters, according to the SPSS algorithm, the optimal number of clusters is three, because the largest ratio of distances is for three clusters. The cluster distribution is shown in Table 3.

Table 3. Cluster distribution

	N	% of Combined	% of Total
Cluster 1	150	15,0%	15,0%
2	351	35,1%	35,1%
3	499	49,9%	49,9%
Combined	1000	100%	100%
Total	1000		100%

SPSS presents also the frequencies for each categorical variable. Table 4 shows the frequencies for *SavingsAccount* variable.

Table 4. Frequencies for *SavingsAccount* variable

	<100		<1000		<500		>=1000		unknown	
	F	%	F	%	F	%	F	%	F	%
Cluster 1	88	14,6%	8	12,7%	16	15,5%	5	10,4%	33	18,0%
2	240	39,8%	20	31,7%	34	33,0%	15	31,3%	42	23,0%
3	275	45,6%	35	55,6%	53	51,5%	28	58,3%	108	59,0%
Combined	603	100,0%	63	100,0%	103	100,0%	48	100,0%	183	100,0%

*Note: * F –Frequencies; % - Percent*

The cluster pie chart from Figure 1 shows the relative size for our three clusters solution.

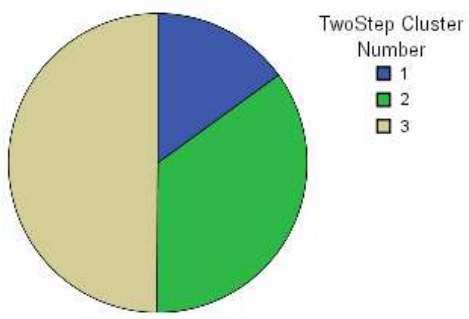

Fig. 1. Cluster size

For categorical variables, the within-cluster percentage plot shows how each variable is split within each cluster. In Figure 2, it is shown the contribution of variable Property within each of the three clusters. Note that in cluster 1, the predominant property is unknown, while in cluster 2 it is the real estate and the car in cluster 3.

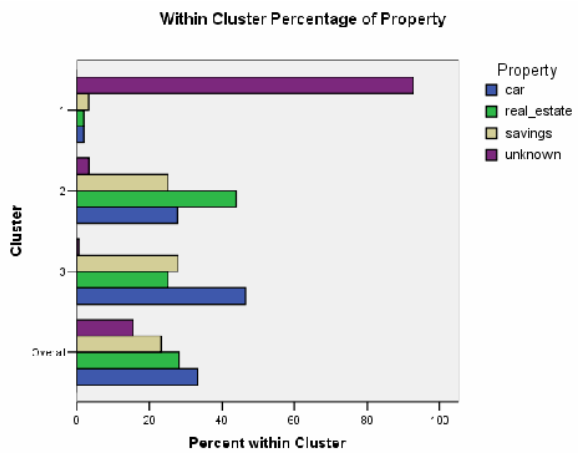

Fig. 2. The weight of *Property* in each cluster

SPSS gives the importance plot for each variable (categorical or continuous). In Figure 3 we present the importance of the categorical variables for the first two clusters.

Note that *Property* and *Housing* contribute the most to differentiating the first cluster and *PersonalStatus*, *CreditHistory*, *Housing* and *YearsEmployed* differentiate the second.

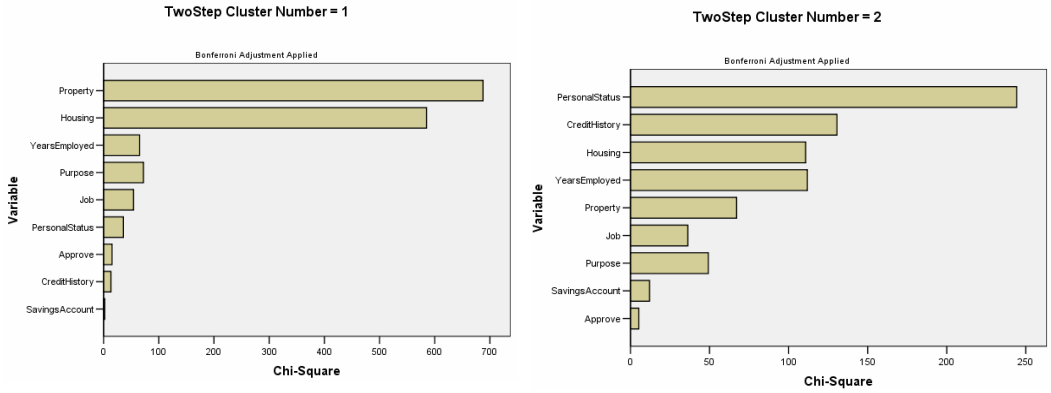

Fig. 3. Categorical variablewise importance for clusters 1 and 2

Regarding the continuous variables importance for cluster 3, we note that cluster 3 is differentiated by the top four variables (the *number of credits at bank*, the *dependents*, the *age* and the *payment rate*) in a positive direction and only by *ResidenceSince* in a negative direction, but the positive variables contribute more to differentiation of cluster 3.

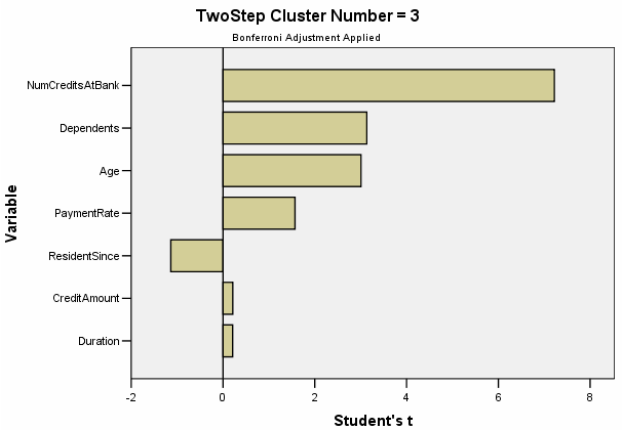

Fig. 4. Continuous variablewise importance for cluster 3

Discussion

We present the following conclusions after the results provided by TwoStep Cluster.

The first cluster, which fills 15%, contains mostly single male customers, which occupy management positions (34.5%) or are unemployed (27.3%), they have unknown properties and their loan is approved in a small percentage (11.9%).

Cluster 2 fills 35.1%, contains female or married male customers with real estate (54.6%), mostly unemployed (54.5%) or unskilled (47.5%) and the purpose of the loan is appliances, retraining and furniture.

The most important cluster is the third. This is the largest cluster (49.9%) containing mostly single male or divorced male customers, with the largest saving accounts, between 4 and 7 years employed, occupying management positions (54.7%) or being skilled workers (50.6%), with a history of credit okay; the purpose of the loan is for business, cars (new or used), or for education; they have their own housing (65.1%) and their loan is approved in a large percentage (55.9%).

Conclusions

Clustering methods can be applied in various fields which use large datasets, just to find hidden patterns. Since most data taken from the real world (as in banking field, in our case) contains both numerical and categorical attributes, classical clustering algorithms can not work efficiently with such data. To solve this problem, we showed that TwoStep method can be easily used, which also determines the optimal number of clusters automatically.

Applying this method to our data, we identified three customers' profiles. The most important profile contains skilled customers with no bad credit history, whose purpose is to obtain the loan for education or for business. The second profile groups middle class customers, unemployed, but with real estate and whose loan is for retraining or for household goods. The third profile groups the persons with unknown properties, mostly unemployed, who want credit for things such as new or used cars or for television, and then for education.

This case study is useful for a bank company which intends to consolidate the company profit, for a better management of existing or possible clients when loan granting.

References

1. *** *Analiza datelor*, available at http://www.spss.ro/detail.php?id=18 , [accessed on 20 July 2010].
2. *** *Euclidian distance*, available at http://en.wiktionary.org/wiki/Euclidean_distance [accessed on 22 July 2010].
3. *** *Source data*, available at http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ [accessed on 9 May 2010].
4. *** *SPSS (Statistical Package for the Social Sciences)*, available at http://www.spss.com/ , [accessed on 10 July 2010].
5. *** *The SPSS TwoStep cluster component*, Technical report, available at http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf [accessed on 20 July 2010].
6. *** *TwoStep Cluster Analysis*, available at http://support.spss.com/productsext/spss/documentation/statistics/algorithms/14.0/twostep_cluster.pdf [accessed 22 July 2010].
7. Arminger, G., Clogg, C., Sobel, M., *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York, 1995, pp. 130.
8. Bacher, J., Wenzig, K., Vogler, M., *SPSS TwoStep Cluster – A First Evaluation*, available at http://www.statisticalinnovations.com/products/twostep.pdf , [accessed on 23 July 2010].
9. Garson, D., *Cluster Analysis*, available at http://faculty.chass.ncsu.edu/garson/PA765/cluster.htm, [accessed 30 July 2010].
10. Ioniță, I., Şchiopu, D., Using principal component analysis in loan granting, *Bulletin of Petroleum-Gas University of Ploieşti*, Mathematics, Informatics, Physics Series, vol. LXI, no. 1/2010.
11. Park, H.-S, Baik, D.-K., A study for control of client value using cluster analysis, *Journal of Network and Computer Applications*, Vol. 29, No. 4, Elsevier, 2006, pp. 262-276.

Aplicarea metodei TwoStep Cluster în identificarea profilului clienților unei bănci

Rezumat

În acest articol se analizează datele referitoare la clienții unei bănci, grupându-i în trei categorii, folosind metoda TwoStep Cluster din SPSS. Această metodă este utilă pentru cazul nostru, deoarece, față de alte metode tradiționale de clustering, TwoStep poate lucra cu date mixte (atât date continue, cât şi numerice) şi poate găsi, de asemenea, numărul optim de clusteri. Metoda creează trei profiluri de clienți. Cel mai mare grup este constituit din clienții calificați, care doresc un credit pentru educație sau pentru afaceri. Al doilea grup conține persoanele care au proprietăți imobiliare, dar care sunt în majoritate şomeri şi care cer un credit pentru a se recalifica sau pentru articole de menaj. Al treilea profil cuprinde persoane cu proprietăți care nu sunt cunoscute şi care cer un împrumut pentru lucruri precum o maşină sau un televizor şi abia apoi pentru educație. Avantajul studiului nostru este consolidarea profitului companiei prin gestionarea mai eficientă a clienților săi.