Clustering Algorithms for Bank Customer Segmentation

Danuta Zakrzewska
Institute of Computer Science,
Technical University of Lodz, Poland
dzakrz@ics.p.lodz.pl

Jan Murlewski
Microelectronics and Comp. Science Dept.,
Technical University of Lodz, Poland
murlewski@dmcs.p.lodz.pl

Abstract

Market segmentation is one of the most important area of knowledge-based marketing. In banks, it is really a challenging task, as data bases are large and multidimensional. In the paper we consider cluster analysis, which is the methodology, the most often applied in this area. We compare clustering algorithms in cases of high dimensionality with noise. We discuss using three algorithms: density based DBSCAN, k-means and based on it two-phase clustering process. We compare algorithms concerning their effectiveness and scalability. Some experiments with exemplary bank data sets are presented.

1. Introduction

The rapid development of data mining methods enables using large data bases of customer data to extract the knowledge, supporting marketing decision process. Areas of knowledge-based marketing, where these techniques are usually applied, are presented in [13]. The wide review of data mining methodology for customer relationship management is given in [11]. As the ability to acquire new clients and retain existing is crucial, especially in the finance marketplace, the possibility of customer segmentation by obtaining the information on unknown hidden patterns has a big significance. Until now only few papers present using of data mining techniques in banks (see [7]). In our work, we consider application of clustering algorithms in this area, taking into account noise possibilities as well as high dimensionality of data.

There are different approaches to cluster analysis for market segmentation. The most popular is k-means algorithm, which together with its modifications was broadly investigated by different authors (see [3], [6], [10], [12]). In a few papers the performance of k-means algorithm was compared with other approaches especially with neural network (see [3],[6],[10]). The last one did not give better results than k-means, but the combination of both was very often recommended as the conclusion.

In our paper we investigate the two-phase clustering algorithm, which consists on modified k-means and hierarchical agglomerative in the second phase. We compare its performance with k-means algorithm and density based approach. As bank customers data sets are usually multidimensional, large and contain noise, we consider effectiveness, scalability and ability of detecting outliers.

2. Clustering algorithms

Cluster analysis techniques have been discussed with details in the literature (see [1], [4],[5],[8],[16]). Efficacy of these techniques depends significantly on characteristic features of data sets. In customer segmentation area, algorithms should be efficient for large multidimensional data sets with noise, what means finding clusters of arbitrary shape and outlier detection. In our research we focus on three algorithms: k-means, which is the most often used method for market segmentation, DBSCAN that performs very good results for large spatial data bases with noise, by forming clusters with arbitrary shape (compare [8]), and the modification of two-phase clustering algorithm that was originally built for outlier detection (see [9]).

2.1. DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is a density based approach to discover clusters and outliers in data bases. The algorithm was introduced by Ester at al. (see [4]) to find arbitrary shape clusters for large spatial databases. The method uses a concept of density-reachability and density-connectivity of points. Cluster points are divided into core points (the ones

Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)
0-7695-2286-06/05 $20.00 © 2005 IEEE

inside the cluster) and border points. The algorithm is based on the following definitions [4]:

Definition 1: A point p is directly density-reachable from a point q with respect to Eps and MinPts if

1. $p \in N_{Eps}(q)$ and
2. $|N_{Eps}(q)| \geq$ MinPts,

where $N_{Eps}(q)$ denotes Eps-neighborhood of point q.

Point p is directly density-reachable from a point q with respect to Eps and MinPts if there exists a chain of points $p_1,\ldots, p_n$, $p_1=q$, $p_n=p$ such that p_{i+1} is directly density-reachable from p_i.

Definition 2: A point p is density-connected from a point q with respect to Eps and MinPts if there is a point o such that both p and q are density reachable from o with respect to Eps and MinPts.

According to these definitions a density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. Every object not contained in any cluster is considered to be noise.

The algorithm starts with any point p and finds out all points density-reachable from p with respect to Eps and MinPts. If p is a core point (Condition 2 in Definition 1) the new cluster is created, if it is a border point, no points are density-reachable from p.

The main advantages of DBSCAN consist on minimal number of input parameters, possibility of discovering clusters of arbitrary shape and outlier detection. In [4], it was also mentioned good efficacy for large data bases, however this property is not fulfilled for high dimensional cases, what will be shown in the next chapter.

2.2. K-means method

In the k-means algorithm, data are partitioned into the given number of clusters. At the beginning, clusters are randomly selected. In each iteration observations are reassigned by moving them into the nearest cluster. New cluster centers are recalculated. The process is continued until all the observations are situated in the closest cluster.

The main advantages of k-means algorithm are its simplicity and effectiveness on large data sets. The main shortcoming is connected with a big dependence of obtained results on the initial assignments, what may entail in not finding the most optimal cluster allocation at the end of the process. Moreover, k-means method is noise sensitive, since even a small number of such observations may influence the result [8].

2.3 Two-phase clustering algorithm

Two-phase clustering algorithm was introduced by M.F. Jiang, S.S. Tseng and C.M. Su in [9], as effective outlier detection technique. It consists of two stages. First phase is the modification of k-means algorithm by using a heuristic "if one new input pattern is far enough away from all clusters, then assign it as a new cluster center". In spite of the original k-means algorithm, where cluster centers are calculated after allocating all the objects, in modified k-means process centers are calculated after every object's allocation. Originally, in the second phase minimum spanning tree is constructed, with clusters obtained in the first phase as nodes, and then the tree is pruned by removing the longest edge. Such procedure allows effective outlier detection.

As the goal in customer segmentation is not only to find out outliers, in the second stage we propose using an agglomerative hierarchical clustering technique. Such combination of algorithms enables to avoid obtaining clusters of low quality, what is the main disadvantage of hierarchical clustering (see[5]). The closest clusters are merged until the desired number of clusters is obtained. The distance between clusters is determined by the single-link approach [8]. Clusters, defined in the first phase, with number of objects below given value are regarded as outliers and are not taken into account in the second stage process. They do not have the influence on the shape of finally determined clusters. Clusters that are not merged with others and contain small number of objects are also marked as outliers.

Let us consider the set of objects $O = \{o_1, o_2, \ldots, o_m\}$, as cluster centers we denote:$C = \{z_1, z_2, \ldots, z_{k'}\}$, where k' means the initial number of clusters. At every iteration the minimal distances between objects and cluster centers $\min(o_i, C)$ are calculated as well as the minimal distance between every two clusters $\min(C) = \min \|z_j - z_h\|^2$, where $j,h = 1, \ldots, k'$; $j \neq h$. Total number of clusters k could be greater then the initial number of clusters k'. To avoid too big amount of clusters, a k_{max} parameter is introduced as the maximal number of clusters. After creating $k_{max}+1$ cluster – the two closest clusters are merged into one.

Modified algorithm has the following steps (compare [9]):

Phase I:

1. Randomly choose k' cluster centers
2. Calculate min(C) and $\min(o_i, C)$ for i=1,2,...,m
3. If $\min(o_i, C) \leq \min(C)$ then go to Step 6

4. Splitting: o_i - center of a new cluster and k'=k'+1
5. Merging: If k' > k_{max} merge two closest clusters and set k'=k_{max}
6. Allocating: assign o_i to its nearest cluster
7. Go to Step 2 until the clusters stabilize.

Phase II:

1. Construct hierarchical tree with clusters as nodes
2. Build the dissimilarity matrix
3. Merge the closest pair of clusters according to dissimilarity matrix
4. Repeat the steps 2-3 until the determined number of clusters is obtained.

The modified algorithm allows to discover clusters of irregular shapes not only spherical, what was easily seen during experiments.

3. Experiments.

In this section we present tests that are performed to compare the algorithms. There were made several experiments on the same data sets with several hundred of records of bank clients data. During tests, algorithms were examined depending on number of dimensions (attributes), efficacy in outlier detection, scalability and behavior in case of standardized and non-standardized data. While testing, the data attributes that are commonly taken into consideration in bank customer analysis were chosen. We consider five values that characterize clients: age, income, deposit, credit, profit/loss. Such choice of attributes enables an identification of groups of customers with similar age, income and credit amounts that give the high profit or the loss to the bank. Discovered segments of clients data will allow specialists of direct marketing to fit their activities to specific target groups.

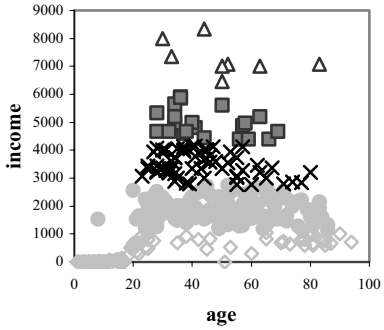

Figure 1. K-means algorithm (k=5).

3.1. Two dimensional case

Figures 1,2 and 3 present the results for all the algorithms in two dimensional case, with attributes reduced to: income and age, on the data set containing 3 hundred records, with noise.

Even in a simple two dimensional case tests showed the differences between efficacy of the algorithms. Two-phase clustering algorithm detected outliers (black squares on Figure 2), while k-means allocated all the objects into clusters.

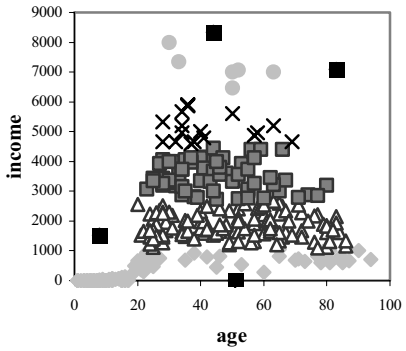

Figure 2. Two-phase clustering algorithm(k'=5)

On the contrary, DBSCAN indicated too many objects as outliers, what we can see on Figure 3. The algorithm presented the tendency to build a small amount of big clusters with many outliers, however its performance depends on the choice of parameters Eps and MinPts.

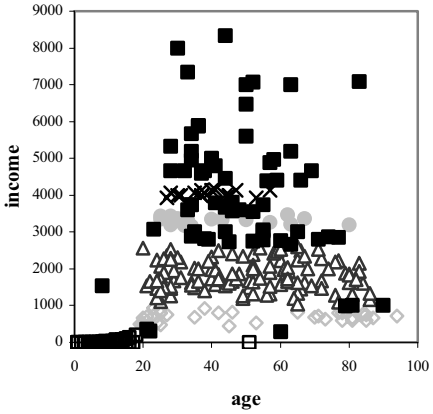

Figure 3. DBSCAN algorithm

3.2. Multidimensional case

As, in the case of bank customers segmentation , it is important to examine the behavior of the algorithms

Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)
0-7695-2286-06/05 $20.00 © 2005 IEEE

depending on the number of dimensions, the next tests were made on the data set with the full range of data attributes: age, income, deposit, credit , profit/loss. In Table 1, 2, 3 we can see cluster centers found by all the algorithms. Tables 4,5,6 shows number of objects allocated in each cluster.

K-means algorithm similarly to the two dimensional case had problems with detecting outliers, and as it is shown in Table 4, it didn't indicate any of them, however the data set contained noise. On the other hand the objects allocation seems to be reasonable even if some clusters contain outliers.

Table 1. K-means – cluster centers

	age	income	deposit	credit	profit/loss
1	29,9	1221,2	1541,0	-1282,7	-14,7
2	42,6	4206,3	14884,9	-108484	316,7
3	58,4	2773,6	44275,7	-2307,7	159,6
4	53,0	5902,3	106622	0,0	379,0
5	46,9	2171,7	9728,0	-417,2	68,0
6	42,5	3013,4	5515,7	-35099	202,3
7	55,6	2386,7	22823,6	-100,3	207,6

Two-phase clustering algorithm has shown the big ability for outliers detection, as it had place in two dimensional case, but the clusters built by the algorithm are not of the same quality. We can observe tendency to create a small amount of big clusters and allocate remaining objects into small clusters (see Table 5).

Table 2. Two-phase clustering algorithm – cluster centers

	age	income	deposit	credit	profit/loss
1	38,7	1727,4	7628,6	-1466,1	40,5
2	61,3	2725,1	37352,6	0,0	188,8
3	53,0	5902,3	106622	0,0	379,0
4	36,1	2391,7	1114,8	-22583	187,6
5	50,5	3491,8	9033,3	-38333	242,0
6	42,6	4206,4	14884,9	-108485	316,7
7	45,8	3393,3	2416,8	-54833	149,0

Table 3. DBSCAN algorithm – cluster centers

	age	income	deposit	credit	profit/loss
1	35,3	1301,7	4859,8	-45,8	21,4
2	58,7	1368,4	21949,9	-30,3	228,6
3	51,4	2961,1	17094,4	-166,6	175,1
4	64,9	1812,4	38809,6	0	73,3
5	37,8	2067,5	373,7	-3789	-70,5

Table 4. K-means - number of objects in clusters

Cluster	Number of objects
1	126
2	11
3	26
4	6
5	70
6	19
7	42

DBSCAN similarly to two dimensional case discovered a big number of outliers and built one big cluster with many objects, as it is presented in Table 6. Other clusters contain only few objects. The same effects were obtained for different input parameters.

Table 5. Two-phase clustering – number of objects in clusters

Cluster	Number of objects
1	236
2	25
3	6
4	9
5	4
6	11
7	4
Outliers	5

Table 6. DBSCAN – number of objects in clusters

Cluster	Number of objects
1	166
2	18
3	8
4	7
5	6
Outliers	95

3.3. Outlier detection ability

Outlier detection ability has a special application in banking sector, particularly in discovering fraudulent operations. There exist several methods dedicated for this goal (see [9]). In customer segmentation, this capacity enables achieving clusters of a good quality, as the process of their creation is not disturbed by noise.

Among the analyzed algorithms, only k-means didn't demonstrate the ability to find out any outliers, while DBSCAN wrongly qualified many objects (see Figure 3 and Table 6). Comparing Figures 1 and 2, we can see that outliers may have a big influence on the shape of clusters in the situation of dispersed structure of objects (the upper clusters on both figures) and is not of such significance in a dense structure case (the lower clusters).

Two-phase clustering algorithms presented good efficacy by indicating real noise as outliers, in both two and multidimensional cases (see Figure 2 and Table 5).

3.4. Scalability

Performance of algorithms were tested on data sets of different size with the same number of attributes. The experiments showed the best performance of k-means algorithm, what might be expected as the effect

of its simplicity. Run time occurred to be directly proportional to the size of data sets. In DBSCAN and two-phase clustering algorithms run times grow much faster then number of objects. In both cases the dependence has an exponential function character. Table 7 shows the comparison of the efficiency of algorithms for three different data sets.

Table 7. Run times in seconds

Number of objects	K-means	Two-phase	DBSCAN
100	0,1563	0,7813	0,1563
500	0,625	3,4688	2,3438
1000	1,25	15,375	9,0625

The two-phase clustering algorithm appeared to be the less scalable, as was expected, because of its complexity. The process of splitting and merging clusters are repeated many times, what every time entails in computing distance matrix . Besides, in the first stage, the algorithm requires higher number of clusters than k-means. Moreover hierarchical method is time consuming for large number of objects.

3.5. Standardization of data

In bank clients data attributes are usually of different character and of different range of values. This feature results in difficulties with choice of a proper similarity function. Standardization allows to classify and to compare such data, by smoothing the differences of data attributes values.

In the two dimensional case the attribute income have much more influence on obtained results, than the attribute age, with its much lower scope of values, what can be easily seen on Figure 1.

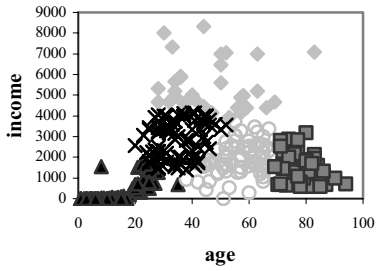

Figure 4. K-means with standardization

Figure 4 presents five clusters built by k-means algorithm with standardization of attributes. Cluster shapes are more regular, because the influence of the attribute age is much bigger, in comparison to test results presented on Figure 1.

The other algorithms show also the good influence of data standardization on the obtained results.

3.6. Choice of parameters

All the considered algorithms were tested for different input parameters by using Euclidean and Manhattan distance functions.

The correctness of results for k-means algorithm depends significantly on the required number of clusters. Experiments showed that in some cases the wrong choice of this parameter leads to completely unsatisfied effects.

Among all the parameters of two-phase clustering algorithm, the demanded number of clusters appeared to be substantial, however the choice of this parameter is intuitive and do not require the knowledge of statistical data distribution. In the tests it didn't cause any problems.

On the contrary to k-means and two-phase clustering algorithms, DBSCAN needs very careful choice of Eps and MinPts parameters, which is rather difficult especially in multidimensional cases. Ester at al in [4] presented the way of finding out parameter values in dependence of determined apriori percentage of outliers, but this approach occurred not to be very easy to implement and effective, particularly for multidimensional cases.

The choice of distance function had no influence on the performance of any of the algorithms.

4. Concluding remarks

In the paper we considered possibilities of applying three algorithms of cluster analysis: k-means, two-phase clustering and DBSCAN in bank customer segmentation. We have taken into account the behavior of algorithms concerning high dimensionality, outlier detection ability and scalability. The tests showed that all the algorithms have their shortcomings and advantages.

K-means algorithm is very efficient for large multidimensional data sets, however depends strongly on the choice of input parameter k. It is not recommended in the case of data sets with noise.

Two-phase clustering algorithm has a very good performance for data with noise and small amount of dimensions. Selection of input parameters is rather easy and intuitive. In multidimensional cases sometimes objects may be allocated into small amount of big clusters.

In DBSCAN algorithm, wrong choice of input parameters, may resulted in a bad quality of obtained

Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)
0-7695-2286-06/05 $20.00 © 2005 IEEE

clusters and the detection of too many outliers. The method becomes effective if Eps and MinPts are determined properly, what is not easy.

Future research will consist on further investigation of the algorithms as well as of the selection of input parameters, and building the modifications that will fulfill the requirements of bank knowledge-based marketing.

5. References

[1] C.C. Aggarwal, C. Procopiuc, J.S. Wolf, P. S. Yu, J. S. Park, "Fast Algorithms for Projected Clustering", *Proc.SIGMOD Conference* ,Philadelphia, 1999, pp. 61-72.

[2] C. Apte, E. Bibelnieks, R. Natarajan, E. Pednault, F. Tipu, D. Campbell, B. Nelson, "Segmentation-Based Modeling for Advanced Targeted Marketing", *Proc. of the XVIIth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Francisco, 2001, pp. 408-413.

[3] P. V. Balakrishnan, M. C. Cooper, V. S. Jacob, P. A. Lewis, "Comparative Performance of the FSCL Neural Net and K-means Algorithm for Market Segmentation", *European Journal of Operational Research,* Elsevier Ltd., 93,1996, pp. 346-357.

[4] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining,* Portland, 1996, pp. 226-231.

[5] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[6] H. Hruschka, M. Natter, "Comparing Performance of Feedforward Neural Nets and K-means for Cluster-Based Market Segmentation", *European Journal of Operational Research,* Elsevier Ltd., 114,1999, pp. 346-353.

[7] N.-Ch. Hsieh, "An Integrated Data Mining and Behavioral Scoring Model for Analyzing Bank Customers", *Expert Systems with Applications,* Elsevier Ltd., 27,2004, pp. 623-633.

[8] A. K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: a Review", *ACM Computing Surveys,*31,3, September 1999, pp. 264-323.

[9] M.F. Jiang, S.S. Tseng, C.M. Su, "Two-Phase Clustering Process for Outliers Detection", *Pattern Recognition Letters,* Elsevier Ltd., 22,2001, pp. 691-700.

[10] R. J. Kuo, L. M. Ho, C. M. Hu, "Cluster Analysis in Industrial Market Segmentation through Artificial Neural Network", *Computers & Industrial Engineering,,* Elsevier Ltd., 42, 2002, pp. 391-399.

[11] Ch. Rygielski, J.-Ch. Wang, D. C. Yen, "Data Mining Techniques for Customer Relationship Management", *Technology in Society*, Elsevier Ltd., 24, 2002, pp. 483-502.

[12] H.W. Shin, S.Y. Sohn, "Segmentation of Stock Trading Customers According to Potential Value", *Expert Systems with Applications,* Elsevier Ltd., 27, 2004, pp. 27-33.

[13] M. J. Shaw, Ch. Subramniam, G. W. Tan, M. E. Welge, "Knowledge Management and Data Mining for Marketing", *Decision Support Systems,* Elsevier Ltd., 31, 2001, pp. 127-137.

[14] C.-Y. Tsai, C.-C. Chiu, "A Purchase-Based Market Segmentation Methodology", *Expert Systems with Applications*, Elsevier Ltd., 27, 2004, pp. 265-276.

[15] P. C. Verhoef, P. N. Spring, J. C. Hoekstra, P. S. H. Leeflang, "The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing in the Netherlands", *Decision Support Systems,* Elsevier Ltd., 34, 2002, pp. 471-481.

[16] M. Zait, H. Messatfa, "A comparative study of clustering methods", *FGCS Journal, Special Issue on Data Mining,* 1997.

Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)
0-7695-2286-06/05 $20.00 © 2005 IEEE