K-means algorithm and its application for clustering companies listed in Zhejiang province

Y. Qian
School of Finance, Zhejiang University of Finance & Economics, People's Republic of China

Abstract

There exist many problems in the credit market where we have data that needs to be classified into distinct groups. This paper will introduce a financial K-means algorithm, which based on the historical financial ratios, applies the cluster analysis technology to analyze the listed enterprises in Zhejiang province. We analyze indicators related to financial attributes and choose nine finance indicators. According to better valuation on the companies listed, we apply "trial and error" and choose four as the number of clustering. Testing shows that companies belong to cluster 2 and cluster 3 add up to 71 companies, including 87% in all. They are all companies worthy of making loans, which is inconsistent with the good economic situation of Zhejiang province. Category 4 has nine companies including 11% that are judged as high risk business. So banks should provide these customers for loans with a mortgage or guarantee.
Keywords: K-means algorithm, clustering analysis, financial ratios, listed companies.

1 Introduction

In credit market, banks want to analyze the customer's preferences to make loan decision, to offer loans and set loan rate, and to decide their market strategy, and to provide customized guide to their potential customers [1].

In today's information based society, there is an urge for bank managers have only vague idea, to find the needed information from the overwhelming resources, who is a good client and who is a bad client (whom to watch carefully to minimize the bank loses. some company's financial reports contain a lot of


WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060041

information that allows us to observe their behavior and operative performance. Properly exploited, this information can assist us to make improvements to loan decision Therefore, data mining, which is referred to as knowledge discovery in database (KDD), has been naturally introduced to the credit market. Clustering plays an important role in knowledge discovery and data mining, and is widely used by banks who need to learn more about their customers. Customers are grouped into clusters by their financial reports, loan granting and setting rate and marketing promotion programs can then be tailored to customers in different clusters.

Traditionally, clustering is done on data sets where an underlying relation R is defined between any two data points in the data set or between any two points in the space containing the data points. The relation between two points is by default considered to be symmetric. However, there are many situations in which the relation is not symmetric (Krishna and Krishnapuram[2]). Guha et al. [3] illustrated that the Euclidean distance can be a poor measure of similarity under this situation. Huang [4] proposed a new distance measure for categorical attributes based on the total mismatches of the categorical attributes of two data records in the k-modes algorithm. In this paper, we introduce financial ratio for clustering attributes. Clustering algorithms may be classified as listed below: Exclusive Clustering; Overlapping Clustering; Hierarchical Clustering, Probabilistic Clustering. However, we propose four of the most used clustering algorithms: K-means; Fuzzy C-means; Hierarchical clustering; Mixture of Gaussians. Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm. We will adopt K-means clustering method in the following paragraphs.

2 K-means algorithm

At a high level, clustering algorithms can be divided into two broad classes: Centroid approaches and Hierarchical approaches. Centroid approaches guess the centroids or central point in each cluster, and assign points to the cluster of their nearest centroid. Hierarchical approaches assume that each point is a cluster by itself, which repeatedly merge nearby clusters, using some measure of how close two clusters are (e.g., distance between their centroids), or how good a cluster the resulting group would be. One widely known unsupervised classification algorithm that is based on clustering the data into local regions is the K-means algorithm (MacQueen [5]). K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem (Kaufman and Rousseeuw [6]; Fukunaga [7]; Duda et al. [8]). K-means is an iterative procedure, including that of Legendre and Legendre [9], to place cluster centers, which quickly converges to a local minimum of its objective function (Bradley and Fayyad [10]; Kanungo et al. [11]).

We have a multivariate input data set X which is defined as an M x N matrix. There are M input points in N-dimensional procedure, including that of Legendre and Legendre [9], to place cluster centers, which quickly converges to a local minimum of its objective function [10, 11]. It is assumed there exist *k* compact classes of data, where $k < n$. The data is classified by allocating each data point to a class and then iteratively moving the data points between classes until we obtain the tightest overall cluster of points in each class.

An algorithm for clustering *N* data points into *K* disjoint subsets containing data points so as to minimize the sum-of-squares criterion

$$J = \sum_{j=1}^{K} \sum_{x \in S_j} | x_n - \mu_j |^2$$

where x_n is a vector representing the nth data point and μ_j is the geometric centroid of the data points in S_j.

The specific algorithm is defined as follows:

Inputs:

$\mathbf{P}=\{p_1 \ldots p_k\}$ *(point to be clustered)*

n *(number of cluster)*

Outputs:

$\mathrm{C}=\{c_1 \ldots c_n\}$ *(cluster centroids)*

$\mathrm{m} : \mathbf{P} \rightarrow \{1 \cdots n\}$ *(cluster membership)*

Procedure K-means $\mathrm{m} : \mathrm{P} \rightarrow \{1 \cdots n\}$

Set C to initial value

For each $p_i \in \mathbf{P}$

$$\mathrm{m}(p_i) = \arg\min_{j \in \{1 \cdots n\}} distance(p_i, c_j)$$

End

Which m has changed

For each $i \in \{1 \ldots n\}$

Recompute c_i *as the centroid of* $\{\mathbf{P}|\mathrm{m(p)}=\mathbf{I}\}$

End

For each $p_i \in \mathbf{P}$

$m(p_i) = \arg\min_{j \in \{1 \ldots n\}} distance(p_i, c_j)$

End

End

End

The algorithm is composed of the following steps:
Step 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
Step2: Assign each object to the group that has the closest centroid.
Step 3: When all objects have been assigned, recalculate the positions of the K centroids.
Step 4: Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The K-means algorithm is unable to handle noisy data and outliers and not suitable to discover clusters with non-convex shapes, and also significantly sensitive to the initial randomly selected cluster centres. In general, the algorithm does not achieve a global minimum of J over the assignments. In fact, since the algorithm uses discrete assignment rather than a set of continuous parameters, the "minimum" it reaches cannot even be properly called a minimum. Despite these limitations, the algorithm is used fairly frequently as a result of its ease of implementation.

3 Samples collecting and data preprocessing

3.1 Companies sampled and financial indicators

The samples of this research come from 81 Securities Issuing in Zhejiang province that have been listed by China Securities Regulatory Commission website and the Great Wall stock trading system [12, 13]. The data draw from the financial data of the third quarter of 2005 on the website. Because the quality of the business enterprise financial standing depends on whether they can pay principal and interest on time or not, we will carry on the classification to these companies to judge its credit risk level and default only by their finance states .so banks make decision on its credit size and lending rate levels and take it reference as developing the credit risk model. Because the finance index sign is numerous, we use SPSS to carry on t test and main compositions analysis to all index, then adopt 9 financial ratio index to measure the business enterprise characteristics:

Liquidity Ratio; Cash Ratio; Equity to Assets Ratio; Inventory Turnover; Assets Liabilities Ratio; Long-Term Liabilities to Assets Ratio; Account Receivable Turnover Rate; Rate of Profit on Net Sales; Return on Total Assets.

3.2 Data preprocessing

Data Preprocessing focuses on two questions: one is how to deal with missing data. For avoiding difficulty in operation the database because of the data defection, this system adopts the average value to deal with missing data.

Another is data standard. As the difference of variables dimension are considered, all data should be standardization. Given some financial ratio x, its

average value is a μ, the standard deviation is a σ, then financial ratio standardized $z=x^*=(x-\mu)/\sigma$, and compute z-scores relative to variables. After data standardize all financial ratio of average value is zero value, its deviation is 1. Thus avoid the question that distance between data points namely between vectors were decide by some value alone.

4 Experimental results

The following table is based SPSS version 14.0. Sample data sets can be found from the official website [12, 13].

This example illustrates application on credit market by the use of K-means clustering with SPSS. The sample data set used for this example is based on the "Zhejiang listed companies data" available in comma-separated format (zjls-data.xls), which includes 81 instances. We will use its implementation of the K-means algorithm to cluster the listed companies in this data set, and to characterize the listed companies segments.

Table 1: Descriptive statistics.

Items	N	Range	Min	Max	Mean	Std. Dev.	Variance
Cash Ratio	81	2.15	0.06	2.21	0.47	0.47	0.22
Equity to Assets Ratio	81	76.02	11.60	87.63	46.59	16.80	283.33
Inventory Turnover	81	75.33	0.01	75.34	6.59	9.44	89.07
Assets Liabilities Ratio	81	0.76	0.12	0.88	0.53	0.17	0.03
Long-Term Liabilities to Assets	81	23.05	0.00	23.05	4.83	5.16	26.67
Rate of Profit on Net Sales	81	66.67	4.10	70.77	21.68	12.50	155.71
Return on Total Assets	81	8.16	0.03	8.19	2.16	1.74	3.01
Valid N (listwise)	81						

Table 1 descriptive Statistics presents some attributes on the real-world instances. Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and normalize the values of attributes that are measured on substantially different scales. SPSS version 14.0 provides filters to accomplish all of these preprocessing tasks. Furthermore, the algorithm normalizes numerical attributes when doing distance computations. The K-means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

Table 2: Initial Cluster Centers.

Zscore	Cluster			
	1	2	3	4
Zscore: Cash Ratio	2.916	-0.863	-0.068	3.389
Zscore: Equity to Assets Ratio	1.866	-0.997	-0.562	2.438
Zscore: Inventory Turnover	7.284	-0.697	0.241	-0.625
Zscore: Assets Liabilities Ratio	-1.863	0.984	0.569	-2.456
Zscore: Long-Term Liabilities to Assets Ratio	-0.910	1.710	-0.612	-0.623
Zscore: Rate of Profit on Net Sales	-0.684	3.934	-0.980	2.698
Zscore: Return on Total Assets	0.471	-0.929	2.620	0.782

Table 2 shows Initial Cluster Centers. The companies listed financial value are used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters. In general, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results. Table 3 Iteration History represents that Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .105. The current iteration is 6. The minimum distance between initial centers is 6.633.

Table 3: Iteration history.

Iteration	Change in Cluster Centers			
	1	2	3	4
1	0.000	2.808	2.779	2.168
2	0.000	0.519	0.106	0.541
3	0.000	0.482	0.165	0.287
4	0.000	0.372	0.177	0.000
5	0.000	0.262	0.201	0.000
6	0.000	0.136	0.101	0.000

Our study is to find each instance along with its assigned cluster. We realign clustering results according to cluster attribute and distance within same cluster (see Table 4 Cluster Membership).

Table 4: Cluster membership.

Case Number	Companies Code	Cluster	Distance	Case Number	Companies Code	Cluster	Distance	Case Number	Companies Code	Cluster	Distance
55	600865	1	0.000	29	600491	2	2.501	74	002019	3	1.713
18	600235	2	0.669	39	600668	2	2.764	5	600097	3	1.729
44	600724	2	0.764	21	600283	2	3.058	48	600796	3	1.775
80	002048	2	1.070	2	600052	2	4.163	49	600797	3	1.776
14	600208	2	1.082	52	600830	3	0.631	57	600982	3	1.870
15	600216	2	1.150	77	002032	3	0.651	22	600330	3	1.896
50	600798	2	1.210	56	600884	3	0.771	40	600671	3	1.896
28	600477	2	1.215	64	000913	3	0.934	35	600571	3	1.938
32	600526	2	1.287	70	002006	3	1.034	66	000963	3	1.965
24	600366	2	1.339	7	600114	3	1.059	54	600857	3	2.168
41	600677	2	1.459	20	600267	3	1.099	36	600572	3	2.236
11	600152	2	1.488	16	600226	3	1.250	58	600987	3	2.305
4	600070	2	1.489	9	600126	3	1.295	65	000925	3	2.345
25	600387	2	1.497	37	600580	3	1.298	19	600261	3	2.407
45	600768	2	1.512	33	600537	3	1.324	60	000517	3	2.675
43	600704	2	1.533	78	002034	3	1.332	10	600130	3	2.786
1	600051	2	1.640	69	002003	3	1.370	51	600814	3	3.176
47	600790	2	1.693	76	002021	3	1.372	38	600596	3	3.195
68	002001	2	1.792	62	000705	3	1.373	72	002011	4	0.552
17	600232	2	1.815	3	600059	3	1.406	6	600113	4	1.180
12	600160	2	1.834	61	000559	3	1.406	67	000967	4	1.632
13	600177	2	1.918	46	600776	3	1.450	59	609036	4	1.931
73	002012	2	1.966	42	600683	3	1.459	23	600340	4	2.223
27	600460	2	1.979	75	002020	3	1.498	81	002050	4	2.298
30	600512	2	1.985	63	000909	3	1.538	34	600570	4	2.461
53	600840	2	2.061	8	600120	3	1.625	31	600521	4	2.464
26	600415	2	2.337	79	002043	3	1.660	71	002010	4	3.561

Table 5 Final Cluster Centers and Table 7 Number of Cases in each Cluster show the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters respectively.

Table 5: Final cluster centers.

	Cluster			
	1	2	3	4
Zscore: Cash Ratio	2.916	-0.478	-0.175	2.067
Zscore: Equity to Assets Ratio	1.866	-0.819	0.151	1.835
Zscore: Inventory Turnover	7.284	-0.220	0.018	-0.161
Zscore: Assets Liabilities Ratio	-1.863	0.820	-0.151	-1.837
Zscore: Long-Term Liabilities to Assets Ratio	-0.910	0.925	-0.479	-0.802
Zscore: Rate of Profit on Net Sales	-0.684	-0.109	-0.097	0.883
Zscore: Return on Total Assets	0.471	-0.474	0.205	0.594

Table 6: Distances between final cluster centers.

Cluster	1	2	3	4
1		9.318	8.296	7.658
2	9.318		2.113	5.068
3	8.296	2.113		3.457
4	7.658	5.068	3.457	

Table 7: Number of cases in each cluster.

Cluster	1	2	3	4
1		9.318	8.296	7.658
2	9.318		2.113	5.068
3	8.296	2.113		3.457
4	7.658	5.068	3.457	

5 Conclusion

As the companies listed are the high-quality customers of banks, it is natural for each bank always to compete. However, because of heavy competition in product market, and difference in the company's management standard level and its strategic marketing. The performance and potentials of the listed companies exist inconsistent. As a result, the banks should classify their customers, before they provide financial services and make a price to carry on the strategy with difference.

K-means algorithm that is introduced in this paper is a kind of clustering methods more maturely, and has application for lot of fields, but some improvements for application are made: we analyze indicators related to financial attributes, and choose 9 finance indicators. According to better

valuation on the companies listed of Zhejiang province, we adopt "trial and error" and choose 4 as the number of clustering.

From experimental results, we find that companies belong to cluster 2 and cluster 3 add up to 71 companies, and including 87% in all. They are all companies worthy of making loans, for banks service they are main customers and profits sources, which is in consistent with good economic situation of Zhejiang province. Category 4 has 9 companies including 11%, compared with other three cluster groups enterprises, they are high risk business and conduct improperly by finance ratios .So the bank should provide these customers for loans with mortgage or guarantee.

However, this study has still some shortcomings: for example the weights have to be determined a priori. Equal weights may result in biased treatment of different attribute types. Moreover, The K-means algorithm has several potential problems including: The classifications depend on the initial values of the class centers chosen. This means sub optimal classifications may be found, requiring multiple runs with different initial conditions.

Acknowledgments

This research is supported by Zhejiang University of Finance & Economics (Grant No.YJZ02), and the project of the National Science Foundation of China (Grant No.70571068).

References

[1] LI Zhihui & LI Meng, Credit risk recognition model and its application for commercial banks in China. *Economic Science, (*5) pp. 61-71, 2005.

[2] K. Krishna & R. Krishnapuram, A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining. *Proceedings of the International Conference on Information and Knowledge Management (CIKM2001)*, Atlanta, Georgia, USA, pp.571-573, 2001.

[3] S. Guha, R. Rastogi & K. Shim, CURE: An efficient clustering algorithm for large databases. *In Proceedings of ACM SIGMOD International Conference on Management of Data.* New York, pp. 73-84, 1998.

[4] Z. Huang, Extension to the k-means algorithm for clustering large data sets with categorical values. Data *Mining and Knowledge Discovery*, 2(3), pp.283-304, 1998.

[5] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1 pp.281-297, 1967.

[6] Kaufman, L. & Rousseeuw, P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, Inc., New York, 1990.

[7] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press Inc., 1990.
[8] Richard O. Duda, Peter E. Hart & David G. Stork, *Pattern Classification.* Second edition, John Wiley & Sons, USA. pp.653, 2001.
[9] Legendre, P. & L. Legendre, *Numerical ecology*. Second edition. Elsevier Science BV, Amsterdam, The Netherlands, 1998.
[10] Bradley & Fayyad, Scaling Clustering Algorithms to Large Databases**.** *Proceedings of the Fifteenth International Conference on Machine Learning* ICML98, pp.91-99, 1998.
[11] T. Kanungo, DM Mount, N. Netanyahu, Christine D. Piatko, Ruth Silverman & Angela Y. Wu, An efficient kmeans clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24, pp.881 -892,2002.
[12] China Securities Regulatory Commission website, http://www.csrc.gov.cn/en/homepage/index_en.jsp.
[13] The Great Wall stock trading system website, www.cc168.com.cn.