CHEMICAL ENGINEERING TRANSACTIONS
VOL. 51, 2016
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian

ISBN 978-88-95608-43-3; **ISSN** 2283-9216
DOI: 10.3303/CET1651085

Financial Distress Prediction of K-means Clustering Based on Genetic Algorithm and Rough Set Theory

Baozhen Hou

Xi'an International University, 18 Yudou Lu, Yanta District, Xi'an Shaanxi, PRC
954212054@qq.com

To the enterprise, the financial risks are impersonal. If there is no scientific method to predict and prevent financial risks, it is likely to cause the enterprise to fall into a difficult situation, even go bankrupt. In reality, the financial crisis of enterprises is expressed as a gradual deterioration of the financial indicators. Therefore, the financial crisis is symptomatic, and can be predicted. The main research direction of the current financial crisis early warning is to looking for a better model to help enterprises to find the financial crisis earlier. In view of the early warning of listed company's financial crisis, this paper introduces the K clustering algorithm based on genetic algorithm; overcomes some problems of traditional K-means clustering, such as sensitive to the initial cluster center, easy to fall into the local optimal value; puts forward a new model which is K-means clustering model based on genetic algorithm. Then, we combine Rough set theory to evaluate financial position of the company comprehensively, and test the rationality of the model classification. Studies have shown that K-means clustering based on genetic algorithm can divide the company efficiently. The goodness of fit to evaluation results of rough set is up to 87.5%.

1. Introduction

For enterprises, if there is no scientific method to predict and prevent financial risks, it will make the enterprise fall into trouble, and even cause bankruptcy. Therefore, it is very necessary to find a sensitive and accurate financial early-warning method. Since 1930s, methods already have more than a dozen through many years development. Over the years, scholars have made the multi-level researches about how to predict the financial crisis of enterprises more accurately. These studies are focused on the following four aspects:

1) The definition of enterprise financial crisis. Altman(1994) defined "entering legal bankruptcy" as the main mark of financial crisis. Deakin(1972) regarded the financial crisis as an enterprise that had been bankrupt or had sold out for insolvency. Foster(1986) treated the financial crisis as a business that must be massive restructured. In our country, Chen Jing(1999) used serious sustained loss as the symbol of financial crisis. Zhao Ailing (2000) took businesses unable to pay due debts or expenses as the symbol of financial crisis. Liu Liang(2006) believed that the financial crisis is defined as the company continued losses. He defined companies suffering losses for two years as a financial crisis company. At present, for company stock in abnormal condition, the Shanghai and Shenzhen Stock Exchange takes special treatments for it. The special treatments include two methods that are delisting risk warning (st*) and other special treatment.

2) Financial crisis early warning model construction. Niu Xiaochen(2014) combined neural network model and rough set theory to predict the financial situation of enterprises. Bao Xinzhong(2013) applied two theories that were K-means algorithm based on particle swarm optimization and rough set theory to predict the financial situation of enterprises. Yang Shue et al. (2005) imposed Back Propagation (BP) neural network model on predict enterprises' financial condition. Lang fan(2008) used improved support vector machine model based on genetic algorithm to predict enterprises' financial condition. Thus, searching for a better algorithm is one of the trends of the current financial distress prediction in various countries. Lai Yuxia(2008) proposed a genetic algorithm based K-means clustering analysis, and improved the original K-means clustering algorithm to construct a better clustering model.

3) Selection of synthetically evaluating indexes. Altman (1968) selected Working Capital / Total Assets, Retained Earnings / Total Assets and other financial indicators in his study. Then, he found that EBIT/ Total

Please cite this article as: Hou B.Z., 2016, Financial distress prediction of k-means clustering based on genetic algorithm and rough set theory, Chemical Engineering Transactions, 51, 505-510 DOI:10.3303/CET1651085

Assets, Sales Revenue / Total Assets, Market Value of Shareholder Equity /Book Value of Total Liabilities showed good predictive ability. In 2006, on the basis of actual situation, Berrada T (2006) selected different methods to find the sensitive indicators by using net operating profit rate and return on equity ratio to construct the early warning model which were obtained good effect. In our country, in the research of financial early warning, song Biao et al. (2015) selected the Liquidity Ratio, Quick Ratio and other 32 financial indicators as the alternative variables. Finally, there are only 5 indicators by using the normal distribution test, such as the Ratio of Fixed Assets, Intangible Assets, etc.

4) Sample selection. Altman et al. (1968) used the data of industrial enterprises Italy between 1982 and 1992 as samples in the study of financial early-warning. In the study, Niu Xiaochen (2014) based on companies listed in the Shanghai and Shenzhen Stock Exchange, and selected the companies which were specially treated from 2010 to 2013.

According to the conclusions above, we define those companies suffered special treatment as the financial distress companies. In addition, we use the K-mean clustering algorithm based on genetic algorithm, which is commonly use in the field of financial early warning. In this paper, we select 28 financial indicators as examine variables, and preserve 7 indicators according to the significant test. We select 24 A shares of listing Corporation as the objects of study. Half of them get special treatment and the others of them are normal financial companies.

2. Algorithm Description

2.1 Architecture of Wireless Sensor Networks

K-means algorithm is one of the most widely used clustering algorithm. The algorithm takes K as parameter, and divides the n objects into K clusters, which makes higher similarity between the intra-cluster, and lower similarity between the inter-clusters. First, the algorithm randomly chooses K objects, each object represents an average or center of a cluster. For the remainder of each object, the algorithm assigns it to the nearest cluster according to the distance from the center of each cluster, then the algorithm recalculates the average value of each cluster. We repeat this progress until the criterion function converges.

The criterion function is as follows:

$$E = \sum_{i=1}^{K} \sum_{X \in C_i} \left| X - \bar{X}_i \right|^2 \tag{1}$$

Among them, $\bar{X}_i$ is the average value of C_i cluster.

K-mean algorithm is described as follows:

(1) K-records are randomly selected as the initial cluster centers.

(2) Calculating the distance between each record and the K cluster center, and regards this point assigned to the clustering which is nearest to it.

(3) Calculating the centroid (mean value) of each aggregation and the distance between each object and the center object, and re-dividing the corresponding object according to the minimum distance. Repeat this step until the formula (1) no longer changes significantly.

Generally, the smaller the value of the formula (1) is, the better the clustering effect is.

2.2 K-mean Clustering Algorithm Based on Genetic Algorithm

The genetic algorithm is applied to cluster analysis. The combination of the global optimal searching ability of genetic algorithm and the local optimization ability of cluster analysis can overcome locality of clustering algorithm. In the process of population evolution, the K-means operation is introduced. Because of the strong local search ability of K, the convergence speed of genetic algorithm can be greatly improved after the introduction of K-means. At the same time, in order to avoid prematurity, the adaptive method is used to adjust the crossover probability and mutation probability dynamically in the population, so that it can be changed automatically with the degree of adaptive function. The specific steps of the algorithm are summarized as follows:

Step 1: Chromosome coding. We code floating point number based on cluster center, and code the center of each category as the chromosome.

Step 2: Generating initial population. In order to obtain the global optimal solution, the initial population is randomly generated. The algorithm randomly assigns each sample to a class as the initial clustering. Then it uses the cluster centers of each cluster as the initial individual chromosome encoded string. Finally, the algorithm generates M initial individuals, so as to generate the first generation of population.

Step 3: Selecting the fitness function. The fitness function is used to measure the individual groups in the optimal solution, in the meantime individuals may reach or close to the optimal solution of the excellent degree. In this paper, we use the formula (1) to construct the fitness function:

f=b/(1+E) (2)

Among them, *b* is constant.

Step 4: Selecting crossover operator, mutation operator, genetic operator. We calculate the probability of individual be selected on the basis of the fitness. Then we replace the worst individual with the current optimal individual in the new population. This article implements the crossover according to the following formula:

$$X_A^{t+1} = aX_A^t + (1-a)X_B^t \ , \ X_B^{t+1} = (1-a)\ X_A^t + aX_B^t \ , \ a \in (0,1) \quad (3)$$

For each aberrance point, algorithm uses a random number in the value of the corresponding gene to replace the original gene and makes the mutation. Algorithm uses the non-fixed probability value for the adaptive genetic operator in the process of genetic algorithm. For the bigger fitness individuals, algorithm gives it corresponding crossover and mutation probability. For example, algorithm sets the corresponding crossover and mutation probability to the individuals which have a bigger fitness. In this way, the adaptive genetic algorithm can keep the diversity of the population while ensuring the convergence of the algorithm.

Step 5: K-means operation. Firstly, we define the values of the new population after the mutation as centered, and assign each data point to the nearest class, so as to form a new cluster. Secondly, we calculate the new cluster center according to the new cluster partition, and replace the original encoding value.

Step 6: Loop termination. Cyclic iteration starts at 0, plus 1 for each cycle. If the current cyclic iteration is less than the pre-specified maximum cycle, then the cycle is continued; otherwise, the end of the cycle. If the current cyclic iteration is less than the predetermined maximum cyclic iteration, the loop continues; otherwise, this loop ends.

2.3 Rough Set

In this paper, rough set is used to attribute reduction of the financial data of the listing Corporation. In the rough set, an information system is defined as *S=(U, C, D, V, f)*. Among them, U is a collection of objects; C and D is collections of attributes; V is a collection of attribute values; and F is an information function, which gives an information value to each attribute of each object in the U. Decision table is a special kind of knowledge expression system. It is a four tuples *T=U, C∪D, V, f*. It means that when certain conditions are met, how decisions should be performed. Definitions are as follows:

Definition 1: The lower approximation of the set X on the R is defined as $R\_(X)=\{x\in U:[x]_R\subseteq X\}$. $R\_(X)$ also known as the R positive domain of X, marked as $POS_R(X)$.

Definition 2: In decision table *S=(U, C, D, V, f)*, the dependency of decision attribute D on conditional attribute set $B\subseteq C$ is defined as $\gamma_B(D) = |POS_B(D)| / |U|$.

Definition 3: In decision table *S=(U, C, D, V, f)*, $c\subseteq C$, the importance of the condition attribute (index) C is defined as $Sig(c) = \gamma_c(D) - \gamma_{C-\{c\}}(D)$.

Definition 4: The weight of the condition attribute (index) C is defined as $W_o(c) = \frac{Sig(c)}{\sum_{\alpha\in C} Sig(\alpha)}$.

In the traditional rough set theory, the attribute is regarded as redundant when the attribute importance is 0. At this point, attribute reduction should be carried out before the calculation of attribute weights.

2.4 Introduction to Financial Early Warning Model

The main body of this paper is based on the genetic algorithm K-means. We use this model to categorize. Besides, we use descriptive statistics and rough set to test results. Step as shown in figure 1:

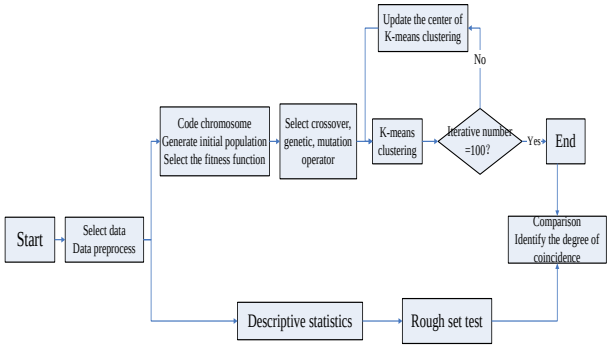

Figure 1: The Principle of Financial Early Warning Model

3 Simulation Experiment and Result Analysis

3.1 Data Selection

If we want predict the monthly sales and get the better results, we need to find out the similar historical month (1) Selection of sample firms. Most of the specially treated company have financial crisis and encounter serious difficulty in its business management. Based on the previous study experience, we define those companies suffered special treatment as the financial distress companies. This paper selects 24 A-share companies listed in Shanghai Stock Exchange and Shenzhen Stock Exchange as the research object. Among them, there are 12 special treatment companies. At the same time, based on selection criteria of similar industry and similar asset size, this paper selects 12 normal financial companies as paired samples. In this paper, the empirical data are derived from “www.cninfo.com.cn/”. (Table 1)

Table 1: Sample Companies

Number (ST*)	Stock symbol	Number	Stock symbol
1	000033	13	600593
2	000059	14	600123
3	600444	15	600060
4	000068	16	300118
5	000155	17	002654
6	000510	18	300446
7	000557	19	300429
8	000590	20	000813
9	000611	21	002144
10	000659	22	002374
11	000677	23	603919
12	000799	24	600519

(2) Selection of financial index. The financial crisis is symptomatic, and can be predicted by a series of observable financial indicators. Whether the financial early-warning indexes are reasonable, it will determine the accuracy of the financial early-warning model. This paper makes full use of previous research in financial warning index selection, initially selects 28 indicators as an alternative to evaluate variables. They are Current Ratio (X1), Quick Ratio (X2), Asset-liability Ratio(X3), Interest Coverage Ratio(X4), Profit Rate to Net Worth(X5), Gross Margins on Sales(X6), Net Profit Margin on Sales(X7), Earning Per Share(EPS) (X8), Receivables Turnover Ratio(X9), Inventory Turnover(X10), Current Assets Turnover(X11), Total Assets Turnover(X12), Net Asset Growth Ratio(X13), Net Profit Growth Ratio(X14), Main Business Income Growth Rate(X15), operating profit growth rate(X16), total profit growth rate(X17), total asset growth(X18), main profit growth rate(X19), cash flow growth rate per share(X20), growth rate of net assets per share(X21), growth rate of earnings per share(X22), ratio of asset inflation proof and incremental value(X23), Ratio of retained earnings to total assets(X24), cash-flow rate(X25), Operating net cash flow per share(X26), fixed assets ratio(X27), and intangible assets ratio(X28).

3.2 Data Classification

We need to classify 24 companies. However, due to the big sample data and the limited space, we only put four companies’ financial data in our paper. These are two st* companies and two normal companies, as shown in the following Table 2 which had pre-processed the defect data.

Table 2: Financial data of companies

Type	Symbol	(st*)000659	(st*)000611	600519	603919
Debt-paying Ability	X1	0.60	2.81	6.15	0.72
	X2	0.47	2.77	4.27	0.38
	X3	67.27	24.55	11.99	58.55
	X4	126.63	-20,436.12	-58,594.82	1,033.92
Profitability	X6	20.25	-4.20	92.61	60.1
	X7	0.41	-141.32	53.01	14.02
	X8	0.00	-0.07	6.91	0.79
Operational capacity	X9	106.96	5,572.76	0.06	3.04
	X10	3.49	2.46	0.11	1.60
	X11	0.93	0.03	0.44	2.03
	X12	0.29	0.02	0.23	0.69
Development ability	X13	-5.43	-10.58	36.54	20.00
	X14	-95.40	-442.98	9.13	32.95
	X15	-20.66	-45.93	10.17	16.74
	X18	-18.05	-4.28	31.00	18.97
Cash flow	X25	10.92	12.35	56.61	31.37
Capital structure	X27	0.59	0.06	0.14	0.37
	X28	0.03	0.07	0.05	0.09

4. Empirical Study and Analysis

4.1 Indicators Screening

Although the primary indicators of financial early warning are very comprehensive, it may not be able to sensitively forecast the Ltd financial situations. Too much of the early warning indexes not only become the heavy workload, but increase the running time of model and affect the accuracy. Therefore, it is necessary for us to further screening of indexes before the construction of early warning model. And there may be a strong correlation between the indicators, which also affect the prediction result.

We use significant tests to screen. Effective early warning indicators should be at least show significant differences between ST* companies and normal companies, so as to correctly judge the financial situation of enterprises. In this paper, the indexes are significant analyzed to complete the screening by the Eviews software. Besides, we remove the indexes which do not contribute to distinguish normal companies from those companies in financial crisis. The results show that only indicators X1, X4, X8, X9, X11, X18, X25 pass the significant test, the rest of indicators are excluded because they don't pass the test of significance.

4.2 Test of Rough

Step 1: Weight calculation. We use the formula of rough set theory to calculate the weight of the 7 financial indexes.

$$W_0(c) = \frac{Sig(c)}{\sum_{\alpha \in C} Sig(\alpha)} \tag{4}$$

Step 2: Comprehensive score for each company. We get comprehensive evaluation scores by $\sum C_i \times W(C_i)$.

Step 3: Test. In order to further test the correctness of clustering results, we use descriptive statistics and variance analysis to analyze the comprehensive score of the companies in each category. The results are shown in table 3.

Table 3: Descriptive statistics situation of the companies

Data	Mean	N	Std. deviation
Class Ⅰ	0.170	8	0.058
Class Ⅱ	0.315	8	0.036
Class Ⅲ	0.412	8	0.072
Total	0.299	24	0.166

4.3 Result analysis

According to the above test, we find that the average is on the upward trend from the first to the third category. It shows that the average score is significant difference between the different categories. The mean errors of the first class, the second class and the third class are small. It shows that different is small within the groups and the companies within each group are similar. There are three companies which are not consistent with the experimental results. We obtain the conclusion that the accuracy of the model is 87.5%.

5. Conclusion

At present, there are a lot of models in the field of financial early warning research, how to choose the appropriate early warning methods play a key role in the validity and accuracy of the warning. In this paper, we find that the K-means clustering algorithm based on genetic algorithm is more accurate than the traditional clustering algorithm. The results of this study provide a new idea for the company classification and a good way of financial warning. It provides an effective basis for the relevant personnel to forecast the financial crisis of enterprises. For protecting the interests of investors and creditors, helping the management departments to monitor listing Corporation quality and avoiding the risk of the securities market, it also has important practical significance.

References

Altman E.I., 1994, Corporate distress diagnosis comparisons using linear discriminate analysis and neural networks (the Italian experience) [J].Journal of Banking and Finance, 524-529.

Altman E.I., 1968, Discriminate Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23, 589-609[J]. Finance, 23(4):589-609.

Bao X.Z., Yang Y., 2013, The Early Warning of Enterprise Financial Crisis Based on Clustering, Rough Set and Neural Network [J]. Journal of Systems & Management, 22(3):358-365.

Berrada T., 2006, Incomplete Information, Heterogeneity, and Asset Pricing [J]. Journal of Financial Econometrics, 4(4):136-160.

Chen J., 1999, The Empirical Analysis on the Prediction of Listing Corporation's Financial Deterioration [J]. Accounting Research, 4:31-38.

Deakin E.B, 1972, A Discriminate Analysis of Prediction of Business Failure [J]. Journal of Accounting Research, 3:167-169.

Foster G., 1986, Financial Statement Analysis [M]. New Jersey: Prentice-Hall.

Lai Y.X., Liu J.P., Yang G.X., 2008, K-Means Clustering Based on Genetic Algorithm [J]. Computer Engineering, 34(20): 200-202.

Lang F., 2008, Research on Financial Early Warning of Improved Support Vector Machine Based on Genetic Algorithm [D]. Journal of Beijing Jiaotong University.

Liu L., 2006, The Empirical Research of Financial Crisis Prediction in Listed Company [D]. Journal of Soochow University.

Niu X.C., 2014, Research on the Financial Early Warning of Listing Corporation which Based on Neural Network Model [J]. Science Technology and Industry, 14(11): 95-98., DOI: 10.1007/978-3-642-25781-0_49

Song B., Zhu J.M., Li X., 2015, Research on Enterprise Financial Early Warning Based on Big Data [J]. Journal of Central University of Finance & Economics, 6: 55-64.

Yang S.E., Huang L., 2005, Financial Crisis Warning Model based on BP Neural Network [J]. System Engineering Theory and Practice, 25(1): 12-18.

Zhao A.L., 2000, Recognition and Analysis of Enterprise's Financial Crisis [J]. The Theory and Practice of Finance and Economics, 21(108): 69-72.