Journal of Computers Vol. 28, No. 1, 2017, pp. 73-86
doi:10.3966/199115592017022801007

A Hybrid Clustering Technique of SM-SOM for Detecting Abnormal Data of Listed Electrical Manufacturing Sector in P. R. China

Rui-cheng Yang[1], Ying Wang[1], and Qing Shen[1]

[1] School of Finance, Inner Mongolia University of Finance and Economics,
No.185, North Second Ring Road, Hohhot, Inner Mongolia, China
yang-ruicheng@163.com, 469375602@qq.com, 1508387502@qq.com

Received 31 August 2015; Revised 17 February 2016; Accepted 10 April 2016

Abstract. For partitioning the dataset of financial ratios into abnormal and normal groups, this paper proposes a hybrid clustering technique by combining similarity matching (SM) algorithm with self-organizing maps (SOM), called SM-SOM technique. The hybrid system provides three stages: preprocessing stage, similarity matching with cosine algorithm and SOM cluster. For evaluating the performance of this hybrid technique, we give some experiments with quarterly financial ratios of listed electrical manufacturing sector in P. R. China. Here the financial ratios contain six categories: profitability, solvency, growth capability, risk level, cash-flow and operating ability, a total of 15 financial ratios are selected such as return on equity, net profit margin, liquidity ratio, and so on. The empirical results show that the SM-SOM technique can improve effectively the accuracy rate for clustering the financial data into normal and abnormal groups. Furthermore, using the hybrid technique we can find out which category these abnormal data fall into.

Keywords: abnormal data, cosine algorithm, electrical manufacturing sector, financial ratio, SM-SOM

1 Introduction

The fraudulent financial reporting of a listed company in stock exchange market can heavily obscure its true credit risk level. This will interfere with the directions of investors and maybe bring great loss to them. So, how to detect the financial fraudulent information of a company has attracted high profile attention of scholars, supervisors and auditors. In fact, financial information of a company can roughly be divided into two classes of abnormal and normal clusters, abnormal data is often regarded as the candidates or red flags of fraudulent financial statements, so, how to partition the dataset of financial ratios into abnormal and normal clusters has become a key issue for detecting the fraudulent financial data of a company. This paper mainly tries to solve this by a hybrid clustering technique, see Section 2 for more details on this.

Clustering algorithms can partition data into a certain number of clusters or groups, and the data in the same cluster are more similar to each other than other different clusters, so, there are considerable interests in the use of clustering techniques to aid in detecting the abnormal data of companies, such as some data mining techniques [1-5] that are dominated by artificial neural networks (ANNs). The popularity of ANNs as a detection tool can be judged from the fact that Green and Choi (1997) [6], Feroz et al. (2000) [7], Spathis et al. (2002) [8], Lin et al. (2003) [9] and Chi-Chen Lin et al. (2015) [10] have used ANNs for detecting the fraudulent financial statement. Recently, one of the most popular and efficient ANNs clustering methods is self-organizing maps (SOM). SOM, proposed by Teuvo Kohonen [11-12], is a type of ANNs that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional) map representation of the input space of the training samples [13]. It is different from other ANNs in the sense that they use a neighborhood function to preserve the topological properties of the

input space. Many researchers used SOM to cluster financial data into different groups (B. Back, K. Sere, H. Vanharanta [14]; Eklund [15]; Jonas Karlsson, Barbro Back, Hannu Vanharanta and Ari Visa [16]). B. Back et al. [14] collected financial statements from 1985 to 1989, which were about 120 multinational companies in North America, North and Middle Europe. They selected nine different financial ratios to verify the SOM and overcame the difficulty of looking for related groups. Eklund et al. [15] took the samples that had been used in the Back's research, evaluated the performance of SOM for the purpose of financial benchmarking of international pulp and paper companies, the results indicated that SOM could be a feasible tool for the financial benchmarking of large amounts of financial data. Dominik Olszewski (2014) [17] proposed a fraud detection method based on SOM. Shin-Ying Huang, Rua-Huan Tsaih and Fang Yu (2014) [18] used growing hierarchical SOM to discover the topological patterns of fraudulent financial reporting. Karlsson et al. [16] analyzed financial performance with quarterly data using SOM, the results showed that the SOM was a feasible and effective tool for financial benchmarking, and were easy to visualize and interpret. SOM can reduce high-dimensional input space to a two-dimensional map and give some visual cluster illustrations. Due to the complexity of financial data, we can still hardly obtain some ideal clustering results. This paper mainly focuses on how to partition the financial dataset into abnormal and normal groups, and the sample data are chosen from the quarterly financial ratios of the listed electrical manufacturing sector in the stock exchange markets of P. R. China. Since the data of each quarterly financial ratio maybe change drastically over time, for instances, the depression of the economy will decline the income sharply of almost all the companies in a same sector, while the related material cost reduction will increase remarkably their respective income. These big fluctuations of financial data will increase the difficulty of fraud identification and cause some risk of misjudgments. However, we find that the quarterly data usually have similar fluctuate features with the other companies which belong to the same sector. So, we introduce the similarity matching method to help us choose the most similar ones as the matched companies, use the data of these matched companies to compute the deviation of the considered company (more details see Section 2), and further derive the sample dataset for SOM clustering. Thus, we get the hybrid technique to cluster the financial data, that is, a combination of similarity matching and SOM (SM-SOM).

Using SOM toolbox in Matlab software 2012b [19], this paper considers the financial ratios which come from six categories: profitability, solvency, growth capability, risk level, Cash-flow and operating ability, a total of 15 indexes. All the financial data are collected from the listed electrical manufacturing sector in the stock exchange market of P. R. China. The empirical results show that the SM-SOM technique can improve effectively the accuracy rate for clustering the financial data into normal and abnormal groups. Furthermore, using the hybid technique we can find out which category these abnormal data fall into.

The paper is organized as follows. Section 2 briefly reviews the hybrid classification model used in this paper, namely, combination of similarity matching (SM) and self-organizing maps (SOM). Section 3 presents some experiments and analyzes the clustering results. Finally, Section 4 gives some conclusions and suggestions for further research.

2 Overview of Methodology

For convenience, we introduce the following notations:

- Considered company A_i $(i=1,2,\cdots,I)$: the i-th considered company that we will cluster its data into abnormal and normal groups.
- Candidate matching company $B_n^i(n=1,2,...,N)$: the n-th candidate matching company for considered company A_i.
- Variable x_{i0}^l $(i=1,2,\cdots,I;\ l=1,2,\cdots,L)$: represents the l-th financial ratio of considered company A_i.
- Variable x_{in}^l $(i=1,2,\cdots,I;n=1,2,\cdots,N;\ l=1,2,\cdots,L)$: represents the l-th financial ratio of candidate matching company B_n^l for considered company A_i.

Now, we give the hybrid classification system in Fig. 1. From Fig. 1, we can see that the architecture of the hybrid classification contains three stages: preprocessing stage, SM stage and SOM stage. More explicit description is given as follows:

- Preprocessing stage: Borrowed the idea of Stigler Stephen M. in 1989 [20], by correlation analysis, we eliminate some high correlation indexes, select L financial ratios as our sample indexes, and further derive the sample dataset of considered company and candidate matching company .
- SM stage: In this stage, we use cosine similarity algorithm to find N matching companies which are the most similar ones to each considered company, and compute the mean value of these matching companies' financial ratios at time t (quarterly time). Based on this, we can obtain the considered companies' deviation data, more details see Section 2.1. These deviation data will be the input data of SOM in the next stage.
- SOM stage: Putting these deviation data of financial ratios into SOM model we can obtain the clustering results: abnormal and normal groups. Furthermore, we can detect which category that the abnormal data of financial ratios fall into.

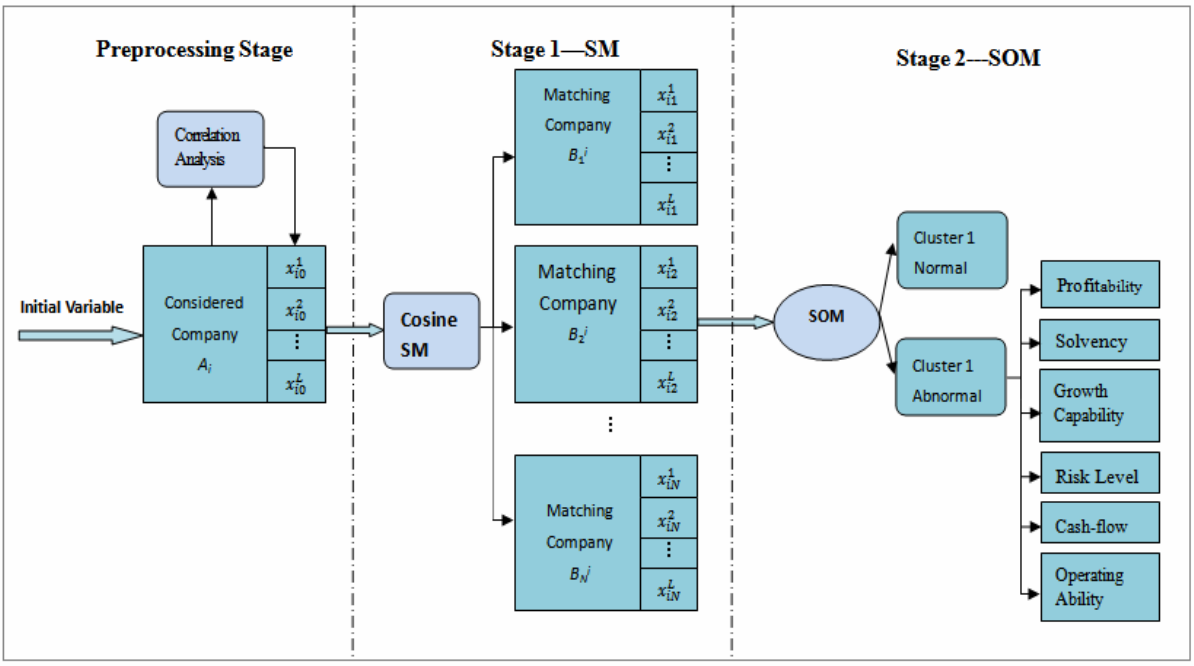

Fig. 1. The architecture of the hybrid clustering technique

2.1 Cosine Similarity Algorithm, Selection of Matching Companies and Deviation Data Computation of Considered Company

Cosine similarity algorithm [20] is to measure the similarity between two vectors by measuring their cosine values of the angle of inner product space. Here, we use the cosine similarity to find some companies that are the most similar ones to the considered company, and further obtain the deviation data of the considered company. Now we give the basic steps as follows:

Step 1—Computation of cosine similarity: Here, we give the cosine similarity S_{in} between considered company A_i $(i=1,2,\cdots,I)$ and candidate matching company B_n $(n=1,2,\cdots,N)$ as follows:

$$S_{in}=\cos(\theta_{in})=\frac{\sum_{l=1}^{L}\sum_{t=1}^{T}x_{i0}^{lt}\times x_{in}^{lt}}{\sqrt{\sum_{l=1}^{L}\sum_{t=1}^{T}\left(x_{i0}^{lt}\right)^2}\sqrt{\sum_{l=1}^{L}\sum_{t=1}^{T}\left(x_{in}^{lt}\right)^2}} \tag{1}$$

where x_{i0}^{lt} and x_{in}^{lt} represent the corresponding quarterly data of financial ratio x_{i0}^{l} and x_{in}^{l} at time $t\in[1,2,\cdots,T]$, and T is the last quarter of resent year that is considered of a company.

Step 2—Selection of matching companies: Now we rank the sequence $\{S_{in}\}$ in descending order with $\hat{S}_1 \geq \hat{S}_2 \geq \cdots \hat{S}_N$, and choose the first several companies as the matching ones. For example, if we choose P $(P \leq N)$ matching companies, we only find the corresponding company by $\hat{S}_1, \hat{S}_2, \cdots \hat{S}_P$. Denote the selected matching companies by ${}_{\mathrm{M}}\mathrm{B}_p$ (p=1, 2, …, P) , and the corresponding ratio data at time t by $\hat{x}_{ip}^{lt}$.

Step 3—Computation of deviation data at time t: Using the data of matching companies ${}_{\mathrm{M}}\mathrm{B}_p$, we give the deviation data of considered company A_i $(i=1,2,\cdots,I)$ as follows:

$$y_i^{lt} = \frac{x_{i0}^{it} - \bar{x}_i^{lt}}{\bar{x}_i^{lt}} \tag{2}$$

where y_i^{lt} is the value of variable y_i^l at time t, $\bar{x}_i^{lt}$ represents the average of the matching companies' data such that

$$\bar{x}_i^{lt} = \frac{1}{P}\sum_{p=1}^{P} \hat{x}_{ip}^{lt} \tag{3}$$

So far, we derive the input data of SOM.

2.2 Self-Organizing Maps (SOM)

Self-organizing map (SOM) network was put forward in 1982 by Kohonen [11]. With the simulation of brain nerve system, the network can realize the function of self-organizing feature map. It is a competitive learning network in which learning can be unsupervised and self-organized.

SOM neural network can map the data of any input mode into one-dimensional or two-dimensional output discrete graphics. It means that the map reduces high dimensional input variables to a low dimensional space, keeping its topology structure unchangeable, and showing the classification results of the data and its correlation with graphic representation. SOM neural network topology structure is divided into input layer and output layer (see Fig. 2). In Fig. 2, the left is input layer that contains the input nodes (input vectors) that each node presents one input variable; the right is output layer which is a two-dimensional plane matrix composed of output neurons. According to certain topological connection and neighborhood functions, each input node is connected with some output neuron nodes of the output layer. The different nodes of the output layer usually represent different classification patterns, of course, there are some exceptions, sometimes, several different nodes maybe represent the same pattern.

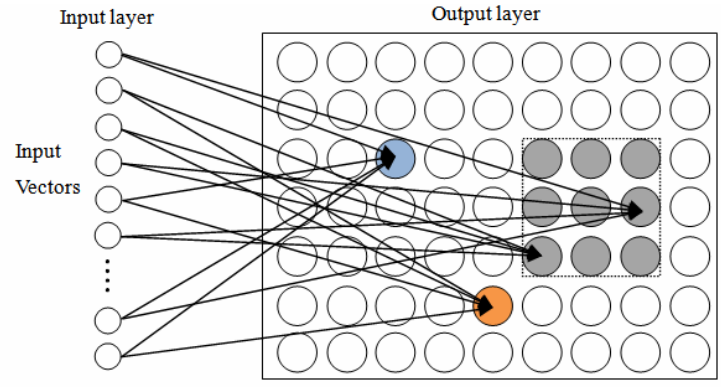

Fig. 2. SOM neural network topology

2.3 Clustering Criterion for Output Topology of Self-Organizing Maps (SOM)

The output layer of a SOM is often presented by U-matrix. Fig. 3 presents a typical topology of U-matrix that is composed of 3×4 neurons. According to the gradually changing color column on the right of the Fig. 3, we can make a decision of the clustering results. For example, in this figure, neuron (3, 3) and neuron (3, 4) with green color mean that the data in these two neurons have similar features, so, they are

regarded as one cluster. Furthermore, we can make a decision that each cluster is abnormal or normal group according to the following decision principles:

- **Majority principle:** it is a binary decision rule that selects alternatives which have a majority, that is, more than half the votes. For instance, in group A with yellow color, if data 1 and data 2 are abnormal, then, we make a decision that the group is abnormal no matter the data 3 is normal or abnormal.
- **Pessimism principle:** This is a conservative decision rule. It puts the safety in the first place. For minimizing the risk as possible, the decision makers prefer to take the worst outcome as the optimal choice. Based on this, when the number of normal and abnormal data in one group is equivalent, the pessimism principle gives us a decision rule that we will regard all these data as abnormal ones. For example, there are data 4-7 in group B with green color, if the true statues of data 4 and data 5 are abnormal while data 6 and 7 are normal, then, we regard all the data in this group as abnormal. In fact, no matter what kind of these data belongs to, it doesn't affect the total accuracy rate. In the real world, we will take further consideration to explore the true reason of the normal data, and correct them.

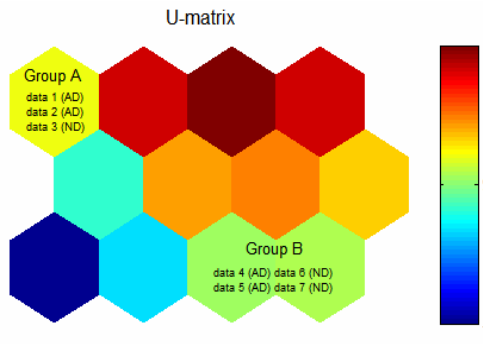

Fig. 3. A typical topology of a SOM U-matrix

3 Empirical Analysis with Hybid Technique

This section gives some experiments which are carried out to evaluate the performance of the proposed hybrid clustering method of SM–SOM.

3.1 Selection of Financial Ratios

Financial ratios of a company are a valuable and easy way to help to answer critical questions such as whether the business is carrying excess debt or inventory, whether the operating expenses are too high, and whether the company assets are being used properly to generate income, and further explain whether the company's profits or other items are reasonable or abnormal [21]. For evaluating the identification performance of the SM-SOM method, we use the quarterly financial ratios of listed electrical manufacturing sector in P. R. China to test it. Based upon some literatures of B. P. Green and J. H. Choi [22], Marilena Mironiuc et al. [23], we take 26 initial financial ratios, and use the correlation analysis to eliminate some strong correlation ratios, finally, 15 explanatory variables are selected as our sample variables. These 15 variables involve the company's six categories of financial ratios: profitability, solvency, growth capability, risk level, cash-flow and operating ability. The definitions and measurements of them are summarized as in Table.1.

3.2 Selection of Companies

In this research, we mainly consider the listed electrical manufacturing listed companies that are chosen from Shanghai and Shenzhen stock exchange markets of P. R. China. Here, the main businesses of these companies mainly concentrate in electrical appliances, machinery and equipment. Due to lack of some financial ratio's data, only 42 companies are entered in our research, of which 3 companies numbered A_1, A_2 and A_3 are involved in financial fraud or suspected fraud and 39 companies numbered B_1, B_2,…, B_{39}

Table 1. Definition and measurement of financial ratios

Category of financial ratios	Definition and Notation	Measurement
Profitability: Profitability measures the company's ability to generate a return on its resources.	Return on equity (ROE) x_{ij}^{1}	Net income/Average equity
	Net profit margin x_{ij}^{2}	Net profit/Primary business income
	Gross profit margin x_{ij}^{3}	Profit/Revenue
	Primary business cost rate x_{ij}^{4}	Primary business cost/Operating costs
Solvency: Solvency measures the ability of a company to meet its long-term fixed expenses and to accomplish long-term expansion and growth.	Liquidity ratio x_{ij}^{5}	Current asset/Current liabilities
	Debt asset ratio x_{ij}^{6}	Total debt/Total asset
Growth capability: Growth ability refers to future development trend and development speed of company, including the expansion of company scale, an increase in profit and owners' equity.	Total Assets Growth Rate x_{ij}^{7}	Asset increment/Total asset
	Net profit growth rate x_{ij}^{8}	Net profit increment/Net profit last year
	Total profit growth rate x_{ij}^{9}	Profit increment/Total profit last year
	Operating profit growth rate x_{ij}^{10}	Operating profit increment/Operating profit last year
	Operating cost growth rate x_{ij}^{11}	Operating cost increment/Operating cost last year
	Revenue growth rate x_{ij}^{12}	Revenue increment/Total revenue last year
Risk level: Risk level is a tool to measure the ability of enterprises to deal with various financial risks.	Financial leverage x_{ij}^{13}	Total debt/Shareholde r's equity
Cash-flow: Cash-flow reflects the number of inflows and outflows of cash for the enterprise's all kinds of activities.	Operating cash-flow rate x_{ij}^{14}	Operating net cash flow/Current liability
Operating ability: Operating ability is used to evaluate the company's utilization degree of their resources and ability of operational activities.	Total assets turnover rate x_{ij}^{15}	Primary business net income/Average total assets

have good credit during the research period, so, the 3 companies are treated as considered companies that contains some normal and abnormal financial data, and the 39 companies are selected as candidate matching companies that only contains normal financial data. We must notice, the sample data's amounts of these companies are different. Since the listing time of company A_1 is later than company A_2 and A_3, so, the data amounts of A_1 is 42 that only covers the period from the first quarter of 2004 to the third quarter of 2014 while other two considered companies are 51 with the period from the Q1/2012 to the Q3/2014, and the periods of considered matching companies change accordingly.

In this paper, the financial data of a considered company are treated as abnormal class mainly includes the following three situations:

- The deviations of the main financial ratios go up and down drastically without any creditable reason.
- If a company's financial statements for a specific period are accused of fraud by P. R. China Securities Regulatory Commission (CSRC), then the data in this period are regarded as abnormal data.
- Fraud record in financial statements.

Compared with other companies, financial data of a considered company appears big difference at the same period.

3.3 Computation of Similarity Matching and Deviation Data

Choosing the matching companies is the key issue for improving the clustering accuracy rate in the hybrid technique. Now, using (1), for each considered company (CC) A_i, we compute its similarity between A_i and candidate matching companies (CMC) B_n^i, and further select the matched companies (MC) with $S_{i\text{-}n}$ (the cosin similarity) $\in [0.65,1]$, here, the first subscript i of $S_{i\text{-}n}$ represents the considered company and the second subscript n represents the matching company, the results are summarized in Table 2.

Table 2. Similarity between CC and MC B_{25}^1

CC	MC	Similarity
A_1	B_{25}^1	$S_{1\text{-}25}$: 0.776
	B_{16}^1	$S_{1\text{-}16}$: 0.760
	B_{21}^1	$S_{1\text{-}21}$: 0.760
	B_{18}^1	$S_{1\text{-}18}$: 0.753
A_2	B_{32}^2	$S_{2\text{-}32}$: 0.837
	B_{33}^2	$S_{2\text{-}33}$: 0.742
A_3	B_3^3	$S_{3\text{-}03}$: 0.823
	B_{16}^3	$S_{3\text{-}16}$: 0.760
	B_{24}^3	$S_{3\text{-}24}$: 0.663

From Table 2, we get four matching companies B_{25}^1, B_{16}^1, B_{21}^1 and B_{18}^1 for considered company A_1, two companies B_{32}^2 and B_{33}^2 for A_2, and three companies B_3^3, B_{16}^3 and B_{24}^3 for A_3. In fact, B_{16}^1 and B_{16}^3 are the same company with the same subscript, the superscript just shows the difference of its corresponding considered company. Next, using (2) again, we can easily get the sample deviation dataset of each considered company A_i (i=1, 2, 3). So far, we derive the input data of SOM.

3.4 Cluster Analysis for Considered Companies

Now we give the clustering results for SOM.

3.4.1 Case for Considered Company A_1

In this section we mainly focus on how to cluster the financial ratio dataset with U-matrix (unified distance matrix) that is generated from the SOM toolbox of Matlab 2012b. The U-matrix is a graphical presentation of SOM. Each entry of the U-matrix corresponds to a neuron in the SOM grid, while value of that entry is the average dissimilarity between the neuron and its neighbors (the values corresponding to colors are depicted in the column on the right).

Now we give the clustering results with the proposed hybrid technique in Fig. 4. It is possible that a hexagon of a given neuron is occupied by more than one data if these data in that hexagon are sufficiently similar. The left of Fig. 4 illustrates the clustering results where the marked letter "ND" in the left of Fig. 4 denotes the normal data while "AD" denotes the abnormal data, and the right of Fig. 4 shows the corresponding exact time of the left figure, for example, the hexagon of a neuron on the top-left of left figure contains "ND2", "ND5" and "ND7", their exact time are "Q2/2014", "Q3/2013" and "Q1/2013" respectively. There are 12 abnormal data and 30 normal data, a total of 42 quarterly data ranged from Q2/2004 to Q3/2014, for convenience, we have arranged these normal data in descending order over quarterly time with 1, 2, …, 30 and abnormal data in descending order over quarterly time with 1, 2, …, 12, such as "AD5" represents the normal data in Q4/2011 while "ND10" represents the abnormal data in Q2/2009. By the majority principle, we get that the column colour range of abnormal

data locates in [0.94, 0.98] ⋃[1.23, ∝), because these hexagons in this colour range include most of abnormal data, and the hexagons with the colour range in (-∝, 0.94) ⋃ (0.98, 1.23) represent most of normal data. Here, we enclose these hexagons of abnormal group with black bold line, and mark the three abnormal data identification errors of "AD6" "AD11" "AD12" and two normal data identification errors of "ND8" "ND27" in red letters. In order to explain more intuitively how to apply the majority principle, let's see the typical hexagon that includes the "AD12, ND16 and ND19", there are two normal data (ND16 and ND19) and one abnormal data (AD12), by the majority principle we classify this hexagon as normal data group with an identification error AD12.

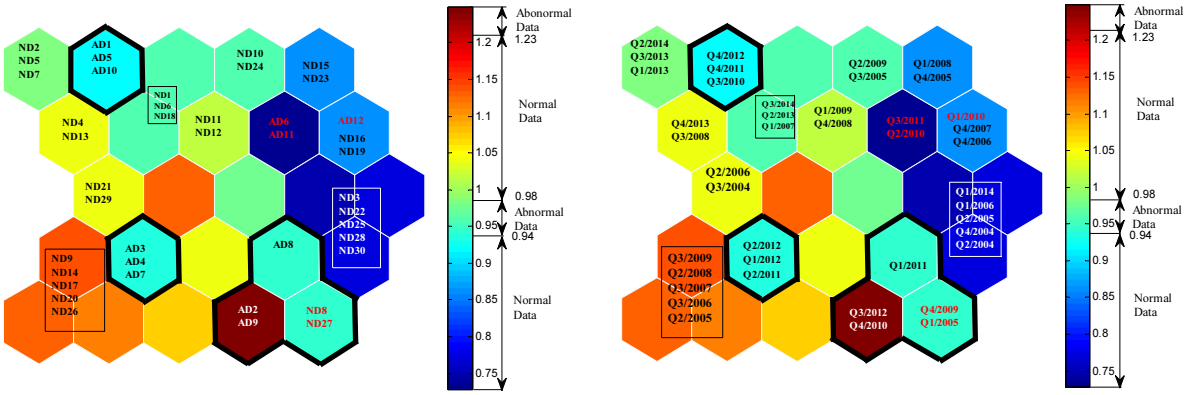

Fig. 4. U-matrix and clustering results of considered A_1 with hybrid technique

It is clear that the SOM visualization of the data can display the majority of the abnormal data. We can see these enclosed abnormal data with the bold line mainly occur from 2010 to 2012. Moreover, for these clusters of abnormal data, we can use the SOM once more to determine which category the abnormal data of financial ratios fall into. From Section 3.1, we know the research financial ratios mainly contain six categories: profitability, solvency, growth capability, risk level, cash-flow and operating ability. Fig. 5 shows the further clustering results of abnormal data, that is, the color of Group A and group B are similar and the categories of this mainly center on profitability and solvency, while group B is coffee color and it occurs mainly on solvency. More explicitly, in group A, the data of Q3/2010, Q1/2011, Q2/2011, Q4/2011, 01/2012, 02/2012 and Q4/2012 are abnormal in solvency and profitability. In group B, the data of Q4/2010 and Q3/2012 are abnormal only in solvency.

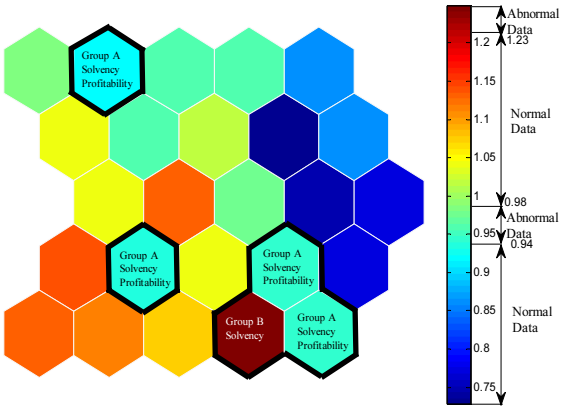

Fig. 5. Categories of company A_1's abnormal financial data

Next, we will give the explanations for the detecting results of abnormal data, and test whether the judgments are in consistence with the true status in the real world. Combined with the financial statements of company A_1 and reports of CSRC, we know that the company A_1 is accused of inflated profits in Q3-Q4/2010, Q1-Q2/2011, Q4/2011 and Q1-Q3/2012, so, these financial data should be regarded as abnormal data that are consistent with the detected results. But there are some exceptions in Q1-Q2/2010 and Q3/2011, we can't find any reason to explain why they are abnormal, then these detected results are

regarded as identification errors.

Table 3. Explanations for detecting results of A_1's abnormal data

Time	Explanations
Q1/2010 Q2/2010	Identification error
Q3/2010 Q4/2010 Q1/2011 Q2/2011	Inflated profits
Q3/2011	Identification error
Q4/2011 Q1/2012 Q2/2012 Q4/2012	Inflated profits

3.4.2 Case for Considered Company A_2 and A_3

Similar to A_1's analysis in Section 3.4.1, Fig. 6 shows the clustering results of considered company A_2, here, the research sample data are ranged from Q1/2002 to Q2/2014, a total of 51 data. The value of the column in Fig. 6 suggests that the abnormal data locates around 0.92. Clustering results show that there are a total of 11 abnormal data and 40 normal data. It is apparent that the enclosed abnormal data are fall into the adjacent hexagons and these hexagons are enclosed with bold line. Among these data, 2 identification errors for real normal data of Q3/2009 and Q3/2012 are determined as abnormal, while Q1/2006 and Q4/2007 that should be real abnormal data are determined as normal, other 47 data with 38 real normal and 9 abnormal data are detected successfully. The detected abnormal data mainly appear in the period from Q1/2006 to Q3/2008.

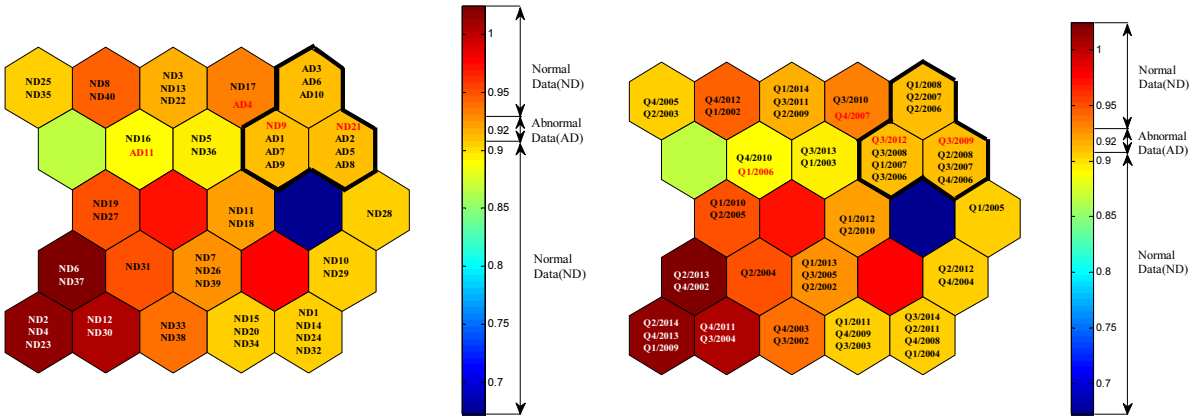

Fig. 6. U-matrix and clustering results of considered A_2 with hybrid technique

Now we explore which category the abnormal data of financial ratios fall into with SOM again. The further clustering results are given in Fig. 7. Similar to the analysis of Fig. 5, we can see that just one abnormal group A consisted of 9 data of Q2-Q4/2006, Q1-Q3/2007, Q1-Q3/2008 is presented (not including two identification errors of data Q3/2009 and Q3/2012 marked with red letters in Fig. 6). And it is composed of three categories of risk level, operating rating ability and cash-flow.

Similar to the analysis in section 3.4.1, we also get the explanations for the detecting results of abnormal data for company A_2, which are shown in Table 4. These explanations contain inflated profits in Q2-Q4/2006, fraud record in incomes, costs and taxes in Q1-Q3/2007, these detected results are consistent with the reports of CSRC. Besides, the financial data in Q1/2008 is treated as abnormal because of its massive growth in profits, although the company explain for this by itself, but its' statements are not reasonable, so it should be treated as abnormal data. In addition, the financial data in Q1/2006 and Q4/2007 are identification errors because there is no any credible reason to explain for this.

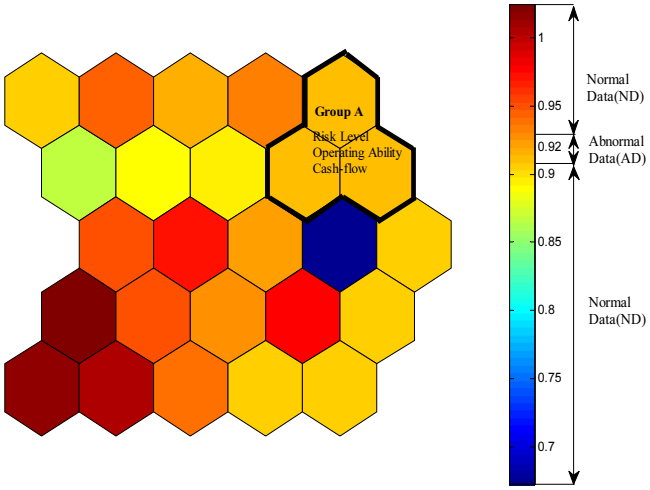

Fig. 7. Categories of company A_2's abnormal financial data

Table 4. Explanations for detecting results of A_2's abnormal data

Time	Explanations
Q1/2006	Identification error
Q2/2006 Q3/2006 Q4/2006	Inflated profits
Q1/2007 Q2/2007 Q3/2007	Fraud record in incomes, costs and taxes
Q4/2007	Identification error
Q1/2008	Massive growth in profits

Similar to the analysis of considered company A_1, Fig. 8 presents that the considered company A_3's colour range of the abnormal data locates in [0.95, 1.38], and the majority centers on the period from Q1/2002 to Q4/2003. Except three identification errors of Q2/2002, Q1/2003 and Q1/2007, there are 9 abnormal data and 42 normal data are classified effectively. The abnormal data are divided into three groups and they are framed in the bold black line. Here, besides the majority principle, we also use the pessimism principle to class one hexagon should be belong to normal or abnormal data group. For example, combining Fig. 8 and Fig. 9 we can see the group A only includes two data, one normal data and one abnormal data, using the pessimism principle we regard the hexagon as an abnormal data group with an identification error. Furthermore, Fig. 9 shows the category that the abnormal data of financial ratios fall into. That is, group A and group B with the data of Q2/2003, Q4/2002 and Q2/2003 are abnormal in cash-flow and risk level, their colour are similar; while group C with light green colour in Q1/2002, Q3/2002 and Q4/2003 are abnormal in all the six categories. And the further explanations for the abnormal data of company A_3 are given in Table 5. From Table 5, we see that the company A_3 was accused of inflated profits and incomes in Q1/2002, Q3-Q4/2002 and Q2-Q4/2003 by CSRC. So, the financial data in these period times are determined as abnormal data. Whereas the data in Q2/2002, Q1/2003 and Q1/2007 are identification errors, they can't be explained reasonably.

3.5 Performance Comparison with Single SOM

For testing the performance of the proposed hybrid SM-SOM method, we derive clustering results for considered A_1, A_2 and A_3 with single SOM. From table 6, we see that the total accuracy rates for all considered companies with SM-SOM are greater than 85%. Comparing with the clustering results of each considered company with single SOM, we know that the performance of SM-SOM is successful. In addition, we also give more intuitive comparisons in Fig. 10 for the identification accuracy rates with respect to normal, abnormal and total data. Fig. 10-a shows that the normal data identification rate of the three considered companies A_1, A_2 and A_3. We can see that the identification rates of company A_1 and A_2 have

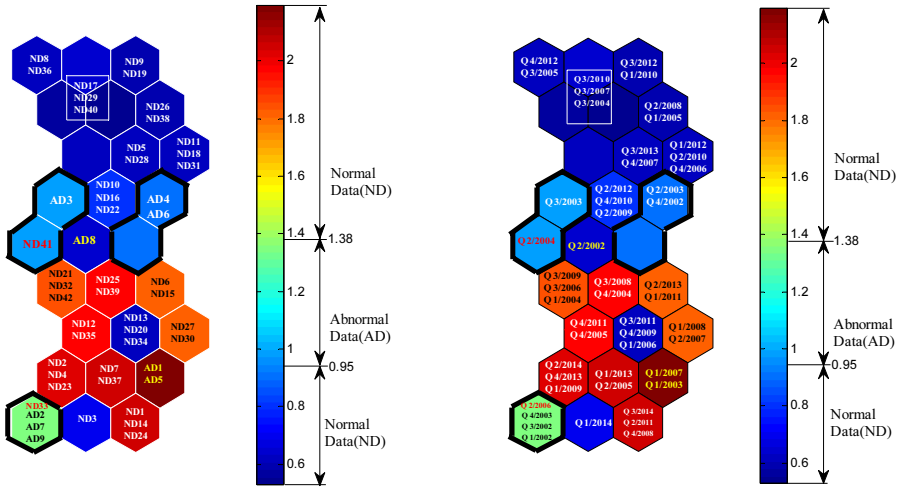

Fig. 8. Categories of company A_3's abnormal financial data

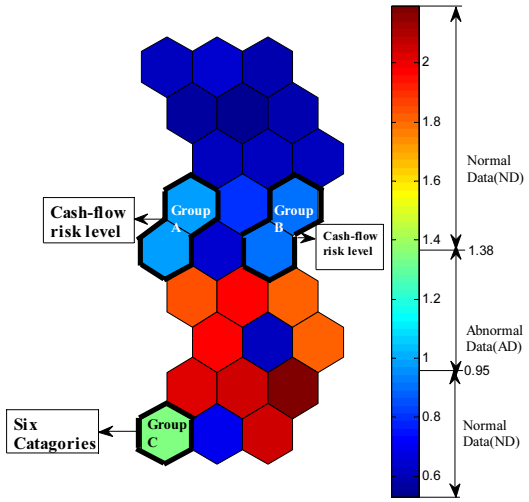

Fig. 9. Categories of company A_3's abnormal financial data

Table 5. Explanations for detecting results of A_3's abnormal data

Time	Explanations
Q1/2002	Inflated profits and incomes
Q2/2002	Identification error
Q3/2002 Q4/2002	Inflated profits and incomes
Q1/2003	Identification error
Q2/2003 Q3/2003 Q4/2003	Inflated profits and incomes
Q1/2007	Identification error

been raised a little while that of company A_3 doesn't improve obviously. However, from Fig. 10-b, for the abnormal data, we can see that all the identification rates of these considered companies improve significantly, the absolute increments are 0.25, 0.4546 and 0.3334 with respect to company A_1, A_2 and A_3, and the corresponding increase proportions are 50%, 125.03%, 100.03%. Accordingly, the total accuracy rates for all the data as illustrated in Fig. 10-c improve obviously. In summary, with the integrated SM-SOM method, the accuracy rate for clustering the financial data has been improved obviously, especially in clustering the abnormal datasets.

Table 6. Statistical results of clustering results with single SOM and SM-SOM

Model	Company	Statue	Total amounts	Number of correct identification	Number of error identification	Accuracy rate (%)	Total accuracy rate (%)
Single SOM	A_1	ND	30	27	3	90.00	78.57
		AD	12	6	6	50.00	
	A_2	ND	40	37	3	92.50	80.39
		AD	11	4	7	36.36	
	A_3	ND	42	40	2	95.24	84.31
		AD	9	3	6	33.33	
SM-SOM	A_1	ND	30	28	2	93.33	88.10
		AD	12	9	3	75.00	
	A_2	ND	40	38	2	95.00	92.16
		AD	11	9	2	81.82	
	A_3	ND	42	40	2	95.24	90.20
		AD	9	6	3	66.67	

4 Concluding Remarks

By combining the SM method with SOM, we propose a hybrid approach for clustering the financial data into normal and abnormal groups. In this SM-SOM system, three stages are provided: preprocessing stage, similarity matching with cosine algorithm and SOM cluster. In preprocessing stage, correlation analysis are used to select the financial ratios; In the SM stage, the cosine similarity method is introduced for selecting matching companies, based on the matching companies, we obtain the deviation of dataset for SOM cluster; Finally, in the last stage, SOM neural network is introduced to cluster the dataset into normal and abnormal groups. For evaluating the performance of this method, we give some experiments with the quarterly financial ratios of listed electrical manufacturing Sector in P. R. China. The empirical results show that the hybrid technique of SM-SOM can improve the accuracy of clustering the financial data into normal and abnormal groups. And furthermore, we can find out easily which category the abnormal data of financial ratios fall into. However, in this research, the classification amounts are still rough, in fact, the abnormal group will be further divided into more detailed clusters for different reasons, this will be our future research work.

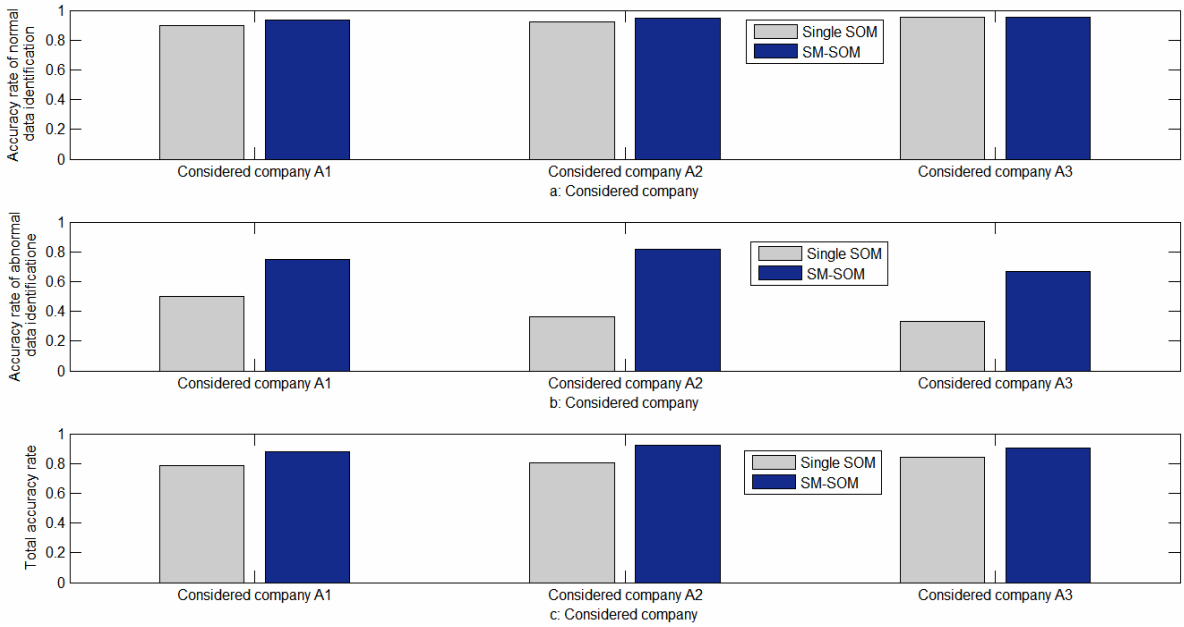

Fig. 10. Accuracy rate comparison of clustering results with single SOM and SM-SOM

Acknowledgements

We thank the referees for their valuable comments, constructive and useful remarks to improve the paper. This work is supported by the National Natural Science Foundation of China (71261015), Program for Innovative Research Team in Universities of Inner Mongolia Autonomous Region (NMGIT1405), Program for Grassland Talent Engineering, Program for Teaching Team in Universities of Inner Mongolia Autonomous Region, and Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region.

References

[1] S. Kotsiantis, E. Koumanakos, D. Tzelepis, V. Tampakas, Forecasting fraudulent financial statements using data mining, International Journal of Computational Intelligence 3(2)(2006) 104-110.

[2] G.N. Singh, G. Rajan, Analysis of data mining techniques for detection of financial statement fraud, The IUP Journal of Systems Management 2012(1)(2012) 7-15.

[3] E. Kirkos, C. Spathis, Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications 32(4)(2007) 995-1003.

[4] E.W.T. Hu, Y. Wong, Y.H. Chen, Y., X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decision Support Systems 50(2010) 559-569.

[5] I. Bose, J. Wang, Data mining for detection of financial statement fraud in Chinese companies, in: Working Paper, The University of Hong Kong, 2008.

[6] B.P. Green, J.H. Choi, Assessing the risk of management fraud through neural network technology, Auditing: A Journal of Practice and Theory 16(1)(1997) 14-28.

[7] E.H. Feroz, M.K. Taek, V.S. Pastena, K. Park, The efficacy of red flags in predicting the sec's targets: an artificial neural networks approach, International Journal of Intelligent Systems in Accounting and Finance Management 9(3)(2000) 145-157.

[8] C. Spathis, M. Doumpos, C. Zopounidis, Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariates statistical techniques, European Accounting Review 11(3)(2002) 509-535.

[9] J.W. Lin, M.I. Hwang, J.D. Becker, A fuzzy neural network for assessing the risk of fraudulent financial reporting, Managerial Auditing Journal 18(8)(2003) 657-665.

[10] C.-C. Lin, A.-A. Chiu, S.-Y. Huang, D. C. Yen, Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, Knowledge-Based Systems 4(9)(2015) 980-989.

[11] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43(1)(1982) 59-69.

[12] T. Kohonen, The self-organizing map, in: Proc. of the IEEE, 1990.

[13] P. Stefanovic, O. Kurasova, Visual analysis of self-organizing maps, Nonlinear Analysis: Modelling and Control 16(4)(2011) 488-504.

[14] B. Back, K. Sere, H. Vanharanta, Managing complexity in large data bases using self-organizing maps, Accounting Management and Information Technologies 8(4)(1998) 191-210.

[15] T. Eklund, B. Back, H. Vanharanta, A. Visa, Data Mining: Opportunities and Challenges, Idea Group Publishing, 2003.

[16] J. Karlsson, B. Back, H. Vanharanta, A. Visa, Analyzing financial performance with quarterly data using self-organizing maps, in: Working Paper, 2001.

[17] D. Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Knowledge-Based Systems70(C)(2014) 324-334.

[18] S.-Y. Huang, R.-H. Tsaih, F. Yu, Topological pattern discovery and feature extraction for fraudulent financial reporting, Expert Systems with Applications 41(9)(2014) 4360-4372.

[19] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhanka nga, Self-organizing map in Matlab: the SOM toolbox, in: Proc. of the Matlab DSP Conference, 1999.

[20] M.S. Stephen, Francis Galton's account of the invention of correlation, Statistical Science 4(2)(1989) 73-79.

[21] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems 50(2)(2011) 491-500.

[22] B.P. Green, J.H. Choi, Assessing the risk of management fraud through neural network technology, Auditing: A Journal of Practice & Theory 16(1)(1997) 14-28.

[23] M. Mironiuc, I.-B. Robu, M.-A. Robu, The fraud auditing: empirical study concerning the identification of the financial dimensions of fraud, Journal of Accounting and Auditing: Research & Practice 2012(1)(2012) 1-13.