J Inf Process Syst, Vol.15, No.3, pp.682~693, June 2019
https://doi.org/10.3745/JIPS.04.0119

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Classification of Imbalanced Data Based on MTS-CBPSO Method: A Case Study of Financial Distress Prediction

Yuping Gu*, Longsheng Cheng**, and Zhipeng Chang***

Abstract

The traditional classification methods mostly assume that the data for class distribution is balanced, while imbalanced data is widely found in the real world. So it is important to solve the problem of classification with imbalanced data. In Mahalanobis-Taguchi system (MTS) algorithm, data classification model is constructed with the reference space and measurement reference scale which is come from a single normal group, and thus it is suitable to handle the imbalanced data problem. In this paper, an improved method of MTS-CBPSO is constructed by introducing the chaotic mapping and binary particle swarm optimization algorithm instead of orthogonal array and signal-to-noise ratio (SNR) to select the valid variables, in which G-means, F-measure, dimensionality reduction are regarded as the classification optimization target. This proposed method is also applied to the financial distress prediction of Chinese listed companies. Compared with the traditional MTS and the common classification methods such as SVM, C4.5, k-NN, it is showed that the MTS-CBPSO method has better result of prediction accuracy and dimensionality reduction.

Keywords

Chaotic Binary Particle Swarm Optimization (CBPSO), Financial Distress Prediction, Mahalanobis-Taguchi System (MTS), Variable Selection

1. Introduction

The classification method of imbalanced data is widely used in various industries, such as medical diagnosis, financial fraud, network intrusion detection, equipment fault diagnosis, etc. The traditional classification methods mostly assume that the class distribution is balanced, and the results obtained by using these methods to solve the problem of imbalanced data are not satisfactory. That is because the amount of information contained in the minority class is easily covered by the majority class, and thus the minority class would be misinterpreted. The minority class should be paid special attention in most imbalanced cases. It is important to research on the imbalanced data issue.

At present, the research of imbalanced data classification mainly focuses on the data field, the algorithm field and the evaluation criterion field [1]. In the data field, it is mainly to reasonably increase the sample

Manuscript received July 3, 2017; first revision September 4, 2017; accepted October 23, 2017.

Corresponding Author: Longsheng Cheng (longsheng_cheng@163.com)
* School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China (gyp1204@163.com)
** School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China (longsheng_cheng@163.com)
*** School of Business, Anhui University of Technology, Maanshan, China (changzp@126.com)

size of minority class or reduce the sample size of majority class. In the algorithm field, it is mainly to include the methods of cost-sensitive learning, integrated learning methods. In the evaluation criteria field, it is mainly to research the evaluation metrics which are including sensitivity, specificity, F-measure, etc. However, these methods of dealing with imbalanced data do not involve variable selection and optimization issues, which often have a significant impact on improving the efficiency of the classification and cannot be ignored.

Mahalanobis-Taguchi system (MTS) [2] is a diagnosis and forecasting method for multivariate data and was proposed by Dr. Taguchi, who is a well-known Japanese quality engineer scientist. MTS is the method based on data analysis, and is easy to understand and implement, so it was widely used in industrial production and fault detection [3-5], enterprise management practice [6-8] and other fields in recent years. Although the MTS method has achieved good results in practical application, there are still some academic controversies. Woodall et al. [9] pointed out that the variable combination with maximum signal-to-noise ratio (SNR) by using the method of orthogonal array may not be the best. Jugulum et al. [10] also mentioned that if a variable optimization method provides a better measurement scale than the orthogonal array and SNR, then this method should be used for the MTS.

The MTS uses a single training sample set of normal class to build the reference space and measure scale, and thus establish a data classification model. The MTS method which uses part of the training samples to learn is very suitable for the classification of imbalanced data. In this paper, the binary particle swarm algorithm combined with chaos theory is used to replace the method of orthogonal array and SNR, and thus proposed the improved MTS method. And for the characteristics of imbalanced data, the evaluation metrics of G-means, F-measure, and dimensionality reduction are selected to improve the classification accuracy and efficiency. Finally, this improved method is applied to the financial distress prediction of Chinese listed companies to show its good performance.

The remainder of this paper is organized as follows. Section 2 presents a short overview of the MTS. Section 3 proposed an improved method for imbalanced data classification. Section 4 applied the proposed method in a case study of financial distress prediction and compared with conventional MTS and other common methods. Section 5 gives some conclusions and future directions.

2. Review of Mahalanobis-Taguchi System

The application of traditional MTS can be carried out in four stages, shown as follows [2].

Stage 1: construct the reference space.

1) Define p variables under healthy or normal conditions.
2) Collect n normal samples, and let x_{ij} is the value of i^{th} sample under j^{th} variable.
3) Normalize the samples of normal group.

$$Z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \qquad i = 1,2,\ldots,n, \qquad j = 1,2,\ldots,p$$
$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad , \quad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2} \tag{1}$$

where, $\overline{x}_j$ is the mean of samples under the j^{th} variable, s_j is the standard deviation of samples under the

j^{th} variable.

4) Calculate the Mahalanobis distance (MD) for all samples in the normal group. The scaled MD of i^{th} sample is used as metric for the reference space and its calculation formula is defined as follows:

$$MD_i = \frac{1}{p} Z'_{ij} C^{-1} Z_{ij} = \frac{1}{p} D_i \tag{2}$$

where C^{-1} is the inverse matrix of correlation matrix, D_i is the standard MD of i^{th} sample.

Stage 2: confirm the validity of the reference space.

1) Identify the unhealthy or abnormal condition. In medical diagnostic application, abnormal conditions are patients with various diseases. In fact, in order to confirm the validity of the reference space, enough abnormal data outside the reference space need to be collected.
2) Normalize abnormal samples by using the mean and standard deviation of normal group, and calculate the MD of abnormal samples according to formula (2).
3) If the MD of abnormal samples are larger than normal samples, the reference space can be judged to be valid.

Stage 3: identify valid variables.

1) In traditional MTS, the valid variables are identified by using orthogonal array and SNR. Select the appropriate orthogonal array according to the number of variables and arrange the variables in the forefront of the orthogonal array. Each variable has two levels, indicating whether the variable is used to build the reference space or not. Each combination of variables in the orthogonal array yields an SNR which is calculated by the MDs of abnormal samples. According to the difference between the two levels of SNR, a valid set of variables can be identified. Larger-the-better SNR is used and its formula is shown as:

$$SNR = -10 \log \left[\frac{1}{t} \sum_{i=1}^{t} \left(\frac{1}{MD_i} \right) \right] \tag{3}$$

where t is the number of abnormal samples.

Stage 4: use valid variables to diagnosis.

1) According to the reference space composed of valid variables, the MDs of unknown samples are calculated. By the comparison between the calculated MD value and threshold value, the various tasks such as classification, diagnosis and prediction can be carried out.

3. Proposed Methodology

3.1 Optimization Goals

3.1.1 Classification performance

For imbalanced data, the overall classification accuracy cannot be simply used as the metric of classification performance, because this tends to predict one sample as the majority class. Therefore, G-mean and F-measure are used as the classification performance metrics. A two-class problem confusion matrix [11] is shown as Table 1, in which true positive (TP) represents the number of positive instance that predicted correctly, true negative (TN) represents the number of negative instances that predicted correctly, false positive (FP) represents the number of positive instance that predicted incorrectly, false negative (FN) represents the number of negative instance that predicted incorrectly. Usually, positive instance belongs to minority class while negative instance belongs to majority class.

Table 1. Confusion matrix of two-class classification

	Prediction positive	Prediction negative
Actual true	True positive (TP)	True negative (TN)
Actual false	False positive (FP)	False negative (FN)

The evaluation metrics of *G*-means and *F*-measure are calculated as follows:

$$G\text{-}means = \sqrt{TPR \cdot TNR} \tag{4}$$

$$F\text{-}measure = \frac{2 * precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

where *Precision*=TP/(TP+FP); *TPR*=TP/(TP+FN), it represents the classification accuracy of positive instance, and is also called Sensitivity or Recall; *TNR*=TN/(TN+FP), it represents the classification accuracy of negative instance, and is also called Specificity.

G-means and F-measure are all composite metric. G-means is the square root of the product of correct rate of two classes, and the larger the classification accuracy of two classes, the larger the value of G-means. Thus G-means can reasonably evaluate the overall classification performance of imbalanced datasets. F-measure is the harmonic mean of the Precision and the Recall, and the larger the Precision and Recall, the larger the F-measure. The value of these two evaluation metrics are both between 0 and 1. The larger the value of the metrics, the better the classification effect of imbalance data.

In general, the normal sample size of MTS is larger than the abnormal sample size. So define the normal sample of MTS as negative instance, and define the abnormal sample of MTS as positive instance. Thus, the parameter of confusion matrix can be obtained by the following method. Let *T* be the threshold of MTS, then:

$$FP = n_{error} = \sum_{i=1}^{n} k_i, \quad k_i = \begin{cases} 1, & \text{if} \quad \mathrm{MD}_i > T \\ 0, & \text{if} \quad \mathrm{MD}_i \le T \end{cases} \qquad i = 1, 2, \ldots, n \tag{6}$$

$$TN = n - n_{error} \tag{7}$$

$$FN = m_{error} = \sum_{j=n+1}^{n+m} k_j, \quad k_j = \begin{cases} 1, & \text{if} \quad \mathrm{MD}_j < T \\ 0, & \text{if} \quad \mathrm{MD}_j \ge T \end{cases} \qquad j = n+1, n+2, \ldots, n+m \tag{8}$$

$$TP = m - m_{error} \tag{9}$$

G-means and F-measure can be calculated as follows:

$$G\text{-}means = \sqrt{\frac{n - n_e}{n} \cdot \frac{m - m_e}{m}} \tag{10}$$

$$F\text{-}measure = \frac{2(m - m_e)}{2m - m_e + n_e} \tag{11}$$

G-means can reasonably evaluate the overall classification performance of imbalanced datasets, and F-measure can properly evaluate the classification performance of minority class. Therefore, these two metrics can be used as larger-the-better optimization goals.

3.1.2 Dimensionality reduction

Another goal of MTS is variable selection and optimization. It is important to identify the variables which are most effective in distinguishing abnormal samples, especially for the high-dimensional data. Dimensionality reduction can improve the classification efficiency and reduce the classification time and cost. Let p_{select} is the number of variables after dimension reduction, and then p_{select}/p represents reduction effect which is another optimization goal of MTS.

3.2 Optimization Mathematical Model

According to the optimization goals, the following optimization mathematical model can be obtained, and is shown as follows.

$$x_i = \begin{cases} 0, & \textit{The variable} \text{ is } \textit{not selected for} \text{ MTS } \textit{analysis} \\ 1, & \textit{The variable} \text{ is } \textit{selected for} \text{ MTS } \textit{analysis} \end{cases} \qquad j = 1, 2, ..., p \tag{12}$$

Then the optimization mathematical model is shown as follows.

$$\min f_1(x_1, x_2, ..., x_p) = 1 - \sqrt{\frac{n - n_e}{n} \cdot \frac{m - m_e}{m}} \tag{13}$$

$$\min f_2(x_1, x_2, ..., x_p) = 1 - \frac{2(m - m_e)}{2m - m_e + n_e} \tag{14}$$

$$\min f_3(x_1, x_2, ..., x_p) = \frac{p_{select}}{p} \tag{15}$$

s.t.

$$f_1(X) < f_1^0 \tag{16}$$

$$f_2(X) < f_2^0 \tag{17}$$

$$\sum_{j=1}^{p} x_j < p \tag{18}$$

$$\sum_{j=1}^{p} x_j = p_{selected} \tag{19}$$

$$x_j = 0 \; or \; 1, \; j = 1, 2, ..., p \tag{20}$$

where f_1^0 equals to 1 minus the G-means of traditional MTS, f_2^0 equals to 1 minus the F-measure of traditional MTS. The constraint conditions (16) and (17) represent the optimized G-means and F-measure should be larger than the traditional MTS method. The constraint conditions (18), (19), and (20) represent whether the variable will be involved, and also represent that the optimized dimensionality reduction effect is superior to the traditional MTS method.

3.3 Optimization Method

For MTS, its variables would either be involved in the construction of the reference space or not, while the threshold T can be a certain range of values. So this is a hybrid constraint nonlinear optimization problem, and binary particle swarm optimization (BPSO) can be used as optimal algorithm [12,13]. In binary search space, the particle swarm consists of N particles in the d-dimensional space, and the position of each particle consists of a string of bits. The position of the particle i is expressed as a vector $X_i=(x_{i1}, x_{i2},\ldots, x_{id})$, where the value of x_{ij} is a binary bit 0 or 1, in which 1 indicates that the corresponding variable is selected and 0 indicates that the corresponding variable is not selected. The velocity of particle i is expressed as a vector $V_i=(v_{i1}, v_{i2},\ldots, v_{id})$. The initial velocity is a random fraction within [0, 1], and the velocity is limited within $[V_{\min}, V_{\max}]$. The vector $P_{besti}=(p_{i1},p_{i2},\ldots,p_{id})$ represents the optimal position that the particle i. The optimal position of all the particles in the population is called the global optimal position, and denoted by $G_{besti}=(g_1,g_2,\ldots,g_d)$. In the BPSO algorithm, the velocity and position of each particle can be calculated according to formula (21) to (24). And then evaluate the optimal position Pb_i and global optimal location Gb_i.

$$v_{ij}^{t+1} = wv_{ij}^{t} + c_1 r_1 (p_{ij} - x_{ij}^{t}) + c_2 r_2 (g_j - x_{ij}^{t}) \tag{21}$$

$$v_{ij}^{t+1} = \begin{cases} V_{\max}, & v_{ij}^{t+1} > V_{\max} \\ V_{\min}, & v_{ij}^{t+1} < V_{\min} \end{cases} \tag{22}$$

$$s(v_{ij}^{t+1}) = \frac{1}{1+e^{-v_{ij}^{t+1}}} \tag{23}$$

$$x_{ij}^{t+1} = \begin{cases} 1, & \xi < s(v_{ij}^{t+1}) \\ 0, & \xi \geq s(v_{ij}^{t+1}) \end{cases} \tag{24}$$

In the above formula, w is the inertia weight, which captures the effect of the previous velocity on the updated one, c_1 and c_2 are acceleration coefficients attached with cognitive and social components of the velocity of a particle, r_1, r_2 and ξ are random numbers in the range [0, 1], v_{ij}^{t} and v_{ij}^{t+1} are the velocity of the particle i before and after the update respectively, x_{ij}^{t} and x_{ij}^{t+1} are the positions of the particle i before and after the update respectively. Eq. (22) shows that the velocity of each particle i is within $[V_{\min}, V_{\max}]$. In order to show that the velocity value is the probability of which the binary bit is taken as 1, the value of the velocity is mapped to the interval [0, 1]. The sigmoid function is usually used as the mapping method, and is shown as Eq. (23) in which $s(v_{ij}^{t+1})$ is the probability that position x_{ij}^{t+1} is taken as 1. Eq. (24) determines the particles to be updated in the next random iteration. If the value of $s(v_{ij}^{t+1})$ is greater than the value that randomly generated within (0,1), then set the value of x_{ij}^{t+1} be 1, otherwise set the value of x_{ij}^{t+1} be 0.

3.4 Chaotic Mapping

In the BPSO algorithm, the inertia weigh controls the global search and local search capabilities of the particle swarm. The large inertia weight is beneficial to the global search, while the small inertia weight is beneficial to the local search. The inertia weight is a key factor influencing the convergence of the problem; it greatly affects the BPSO search process, and thus affects the accuracy of prediction. Because the BPSO algorithm is easy to fall into the local optimal and this would lead to convergence early, chaotic mapping is introduced into the BPSO algorithm to form a chaotic BPSO (CBPSO) algorithm [14], which could overcome the shortcoming of premature convergence.

Chaos is the method that the non-deterministic stochastic state can be obtained from deterministic equation. Chaos has the characteristics of randomness, ergodicity and regularity. So in the each iteration of the BPSO algorithm, the chaotic map is used to determine the inertia weight. The values of the inertia weights are usually calculated using logistic maps.

$$w(t+1) = uw(t)(1-w(t)) \qquad t = 0,1,2... \tag{25}$$

where $w(t)$ is a chaotic sequence and its value is limited within (0,1); u is the control parameter, and when $3.571448 \leq u \leq 4$, the logistic map is in a chaotic state, especially when u=4, it is in a completely chaotic state. When the inertia weight is close to 1, the global optimal search ability is enhanced. When the inertia weight is close to 0, the local optimal search capability is enhanced.

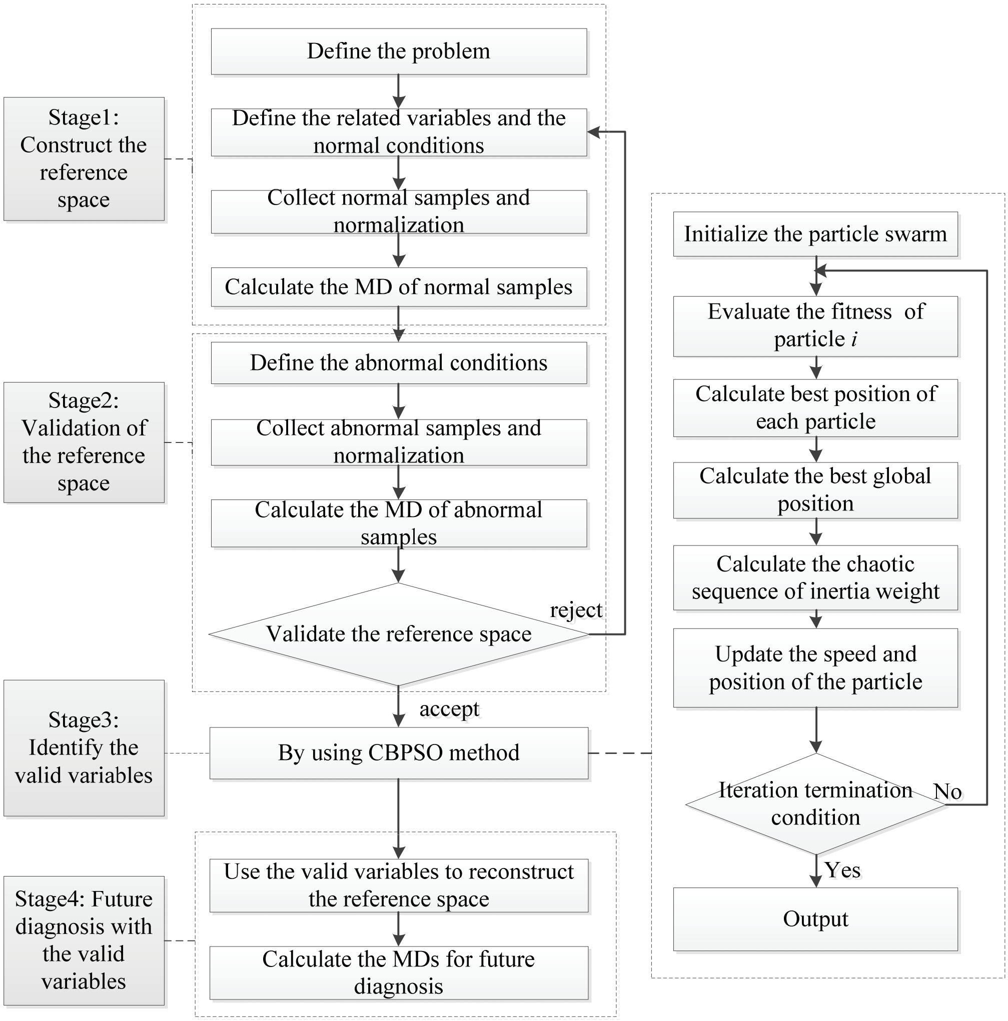

Stage1: Construct the reference space
Define the problem
Define the related variables and the normal conditions
Collect normal samples and normalization
Calculate the MD of normal samples
Stage2: Validation of the reference space
Define the abnormal conditions
Collect abnormal samples and normalization
Calculate the MD of abnormal samples
Validate the reference space
reject
accept
Stage3: Identify the valid variables
By using CBPSO method
Stage4: Future diagnosis with the valid variables
Use the valid variables to reconstruct the reference space
Calculate the MDs for future diagnosis
Initialize the particle swarm
Evaluate the fitness of particle i
Calculate best position of each particle
Calculate the best global position
Calculate the chaotic sequence of inertia weight
Update the speed and position of the particle
Iteration termination condition
No
Yes
Output

Fig. 1. The flow chart of MTS-CBPSO method.

3.5 MTS-CBPSO Method

According to the optimization mathematical model, the multi-objective optimization problem could be transformed into a single-objective problem by integrating all optimization goals. Let the fitness function of BPSO be $Min\ f_1\ f_2\ f_3$. The basic idea of MTS-CBPSO method is: N particles are produced in each time when the CBPSO algorithm iterates; according to the position of each particle, select the corresponding combination of variables that participate in the operation, then the training samples are classified by MTS method by using the combination of these variables, and compare the fitness values of all particles to obtain the current optimal particle; the optimal fitness value and its corresponding variable combination could be obtained through iteration of the CBPSO algorithm. The variables combination can be regarded as the output of BPSO, and then implement the test sample according to the traditional MTS method.

Similar to the traditional MTS, MTS-CBPO method can be divided into four stages as shown in Fig. 1.

4. Case Study: Financial Distress Prediction

In an economic globalization environment, because of the intense market competition, enterprises are facing various risks during the process of development, especially financial risk. The financial distress can reflect the increased financial risk. So effective financial distress prediction is not only related to the development of the enterprise itself, but also related to the interests of investors, creditors and other stakeholders. Therefore, the study of financial distress of listed companies is a hot topic in the field of corporate governance, risk control and securities investment research in capital market. At present, the financial distress forecast of listed companies is basically regarded as a classification problem. That is, according to the financial situation, enterprises will be divided into two categories of being normal or abnormal. But these studies usually artificially balance the normal and abnormal samples, while ignoring the nature of their own imbalance [15,16]. The following will describe how CBPSO-MTS algorithm is used to study and analyze the financial distress of listed companies.

4.1 Choice of Samples and Variables

Take Chinese listed companies as the research objects. Companies that are specially treated (ST) by the China Securities Supervision and Management Committee (CSSMC) are considered as companies in financial distress and those who are never ST are regarded as healthy ones. According to the data between 2010 and 2015, select 150 listed companies which were ST due to abnormal financial situation as abnormal sample, and select 350 listed companies which were never ST as normal sample companies of healthy financial status. The data used in this study was obtained from RESSET database (www.resset.cn). In order to eliminate outliers, companies with financial ratios deviating from the mean value as much as three times of standard deviation are excluded, and thus eventually get 425 sample companies, among which 115 are ST companies and 310 are normal ones.

Because different companies have different reasons to be treated as ST, it is difficult to use simple financial ratio metrics to describe the company's financial situation. And different researchers choose the different financial ratio metrics. To truly reflect the financial situation of enterprises, 52 metrics of primary financial ratio are selected, and then remove the metrics that have a correlation coefficient greater than 0.95, remaining 38 indicators which can be included in six dimensions such as profitability, solvency, business development capacity, operational capacity, cash flow and capital structure, as shown in Table 2.

Table 2. Financial ratio indicators system

Dimensions	Indicators name
Profitability	profit margin on net assets X_1; return on assets ratio X_2; net profit to total assets X_3; net profit to total operation income X_4; total operation cost to total operation income X_5; asset impairment loss to total operation income X_6; operating profit ratio X_7; total profit cost ratio X_8
Solvency	current ratio X_9; super quick ratio X_{10}; debt to equity ratio X_{11}; earnings before interest to total liability X_{12}; net operating cash flow to total liability X_{13}; net operating cash flow to total current liability X_{14}; cash flow to liability X_{15}
Business development capacity	earnings per share growth rate X_{16}; operating profit growth rate X_{17}; total profit growth rate X_{18}; net profit growth rate X_{19}; net operation cash flow growth rate X_{20}; net asset growth rate X_{21}; total assets growth rate X_{22}
Operational capacity	inventory turnover X_{23}; receivables turnover ratio X_{24}; account payable turnover rate X_{25}; current assets turnover rate X_{26}; fixed asset turnover rate X_{27}; total asset turnover rate X_{28}
Cash flow	sales and service cash to operating income X_{29}; capital expenditure to depreciation and amortization X_{30}; operating income cash coverage X_{31}; operation cash into Asset rate X_{32}
Capital structure	debt to asset ratio X_{33}; current asset to total asset X_{34}; fixed asset ratio X_{35}; equity to total capital X_{36}; current liability to total liability X_{37}; long asset fit asset X_{38}

4.2 Experiments and Results

The positive instance refers to the sample of minority class, which is the ST company in this case, and is the abnormal sample of MTS. The corresponding negative instance refers to the non-ST company, which is the normal sample of MTS. The experiment uses the 5-fold cross-validation method, that is, each time the data sets are randomly divided into five parts, each experiment selected four of them randomly as training samples, and the remaining one is used as validation sample. All experiments were performed using the *rminer* packages and *R* tool. The classification ability is evaluated by the mean of the results of the five cross experiments. The evaluation criteria are: sensitivity, which is the classification accuracy rate of abnormal sample; specificity, which is the classification accuracy rate of normal sample; accuracy, which are the total classification accuracy rate; G-means and F-measure.

The parameters of CBPSO algorithm are set as the classic value, that is, c_1=c_2=2; the number of particles N=30; the lower and upper limits of the particle velocity are -2 and 2 respectively; and the maximum number of iterations is 150. The initial value of chaotic sequence $w(0)$=0.48, u=4, which can ensure that the chaotic system is completely in chaotic state. The convergence of CBPSO is demonstrated in Fig. 2.

The 5-fold cross-validation experiments were carried out using the traditional MTS and MTS-CBPSO, respectively. The result is shown in Table 3.

In terms of variable optimization, the average number of variables being deleted in MTS is 8.6, while the average number of variables being deleted in MTS-CBPSO is 14.8. This indicates that MTS-CBPSO has a better effect of dimensionality reduction. It is because that the evaluation metrics are appropriately set for the imbalanced data, and when the method iterative to achieve optimal, the time cost which is represented by the number of variables is also taken into account. So as to achieve the same classification accuracy, the effect of dimensionality reduction is also considered.

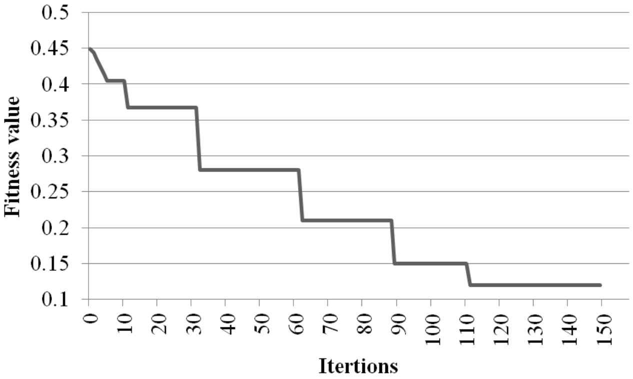


Fig. 2. Convergence of CBPSO algorithm.

Table 3. Comparison of classification effects

	Sensitivity	Specificity	Accuracy	G-means	F-measure
MTS-CBPSO	0.896	0.915	0.909	0.902	0.908
MTS	0.824	0.845	0.832	0.838	0.831

In order to further explain the effect of the MTS-CBPSO method on the financial distress forecast of listed companies, the algorithms of SVM, C4.5 and *k*-NN (*k*=5) are carried out with the same data. Five-fold cross-validation method is also used, and all metrics take the average value of five experiments. The result is shown in Table 4.

Table 4. Comparison of other algorithm classification effects

	Sensitivity	Specificity	Accuracy	G-means	F-measure
SVM	0.819	0.849	0.830	0.835	0.826
C4.5	0.788	0.829	0.812	0.812	0.801
k-NN	0.716	0.842	0.788	0.772	0.754

4.3 Discussion

From the above analysis, it is obviously that the MTS-CBPSO method is superior to the traditional MTS in both dimensionality and classification accuracy. What's more, the MTS-CBPSO is more robust. Compared with the other three algorithms, the classification result of the normal sample does not have much difference. But as to the classification result of the abnormal sample, MTS-CBPSO is best, followed by are SVM and C4.5, *k*-NN is worst. The classification result of *k*-NN method is acceptable with majority class, but is worst with minority class, and this is due to the relatively small amount of abnormal financial class and the relatively large noise, while in the financial distress prediction process, minority class sample recognition is more important. Therefore, the MTS-CBPSO algorithm has the best effect on the financial distress prediction of listed companies; MTS, SVM, C4.5, and *k*-NN are relatively weak.

5. Conclusions and Future Directions

Compare with previous study, the contribution of this paper is showed as follows. (1) Based on the traditional MTS, the BPSO algorithm is used to replace the method of orthogonal array and SNR to optimize the variables, and furthermore, the chaotic mapping is introduced into the BPSO algorithm. Because of the fast convergence of BPSO algorithm and global ergodic characteristics of chaos mapping, BPSO algorithm could effectively get rid of the local convergence value, and so that the optimization accuracy could be improved and stable. (2) According to the characteristics of imbalanced data, G-means, F-measure, and dimensionality reduction are used as classification metrics instead of the overall correct rate of classification. (3) Apply CBPSO-MTS method to predict the financial distress of Chinese listed companies based on 38 financial metrics. The results show that the performance of MTS-CBPSO is better than traditional MTS and other commonly used classification methods, and is more suitable to deal with imbalanced data.

However, the financial data used are derived from the annual report, while the data of quarterly and semi-annual report are not used. These data may be more timeliness for financial distress prediction, which are the follow-up research directions.

Acknowledgement

This paper is partially supported by National Natural Science Foundation of China (Grant No. 71271114 and 71673001).

References

[1] S. Maldonado and J. Lopez, “Imbalanced data classification using second-order cone programming support vector machines,” *Pattern Recognition*, vol. 47, no. 5, pp. 2070-2079, 2014.

[2] G. Taguchi and R. Jugulum, *The Mahalanobis-Taguchi strategy: A Pattern Technology System*. New York, NY: John Wiley & Sons, 2002.

[3] P. Shakya, M. S. Kulkarni, and A. K. Darpe, “Bearing diagnosis based on Mahalanobis–Taguchi–Gram–Schmidt method,” *Journal of Sound and Vibration*, vol. 337, pp. 342-362, 2015.

[4] B. John, “Application of Mahalanobis-Taguchi system and design of experiments to reduce the field failures of splined shafts,” *International Journal of Quality & Reliability Management*, vol. 31, no. 6, pp. 681-697, 2014.

[5] X. Jin and T. W. Chow, “Anomaly detection of cooling fan and fault classification of induction motor using Mahalanobis–Taguchi system,” *Expert Systems with Applications*, vol. 40, no. 15, pp. 5787-5795, 2013.

[6] B. Valarmathi and V. Palanisamy, “Opinion mining of customer reviews using Mahalanobis-Taguchi system,” *European Journal of Scientific Research*, vol. 62, no. 1, pp. 95-100, 2011.

[7] S. E. Abbasi, A. Aaghaie, and M. Fazlali, “Applying Mahalanobis–Tagouchi system in detection of high risk customers: a case-based study in an insurance company,” *Journal of Industrial Engineering*, vol. 45, no. 2, pp. 1-12, 2011.

[8] C. L. Huang, Y. H. Chen, and T. L. J. Wan, “The Mahalanobis Taguchi system: adaptive resonance theory neural network algorithm for dynamic product designs,” *Journal of Information and Optimization Sciences*, vol. 33, no. 6, pp. 623-635, 2012.

[9] W. H. Woodall, R. Koudelik, K. L. Tsui, S. B. Kim, Z. G. Stoumbos, and C. P. Carvounis, "A review and analysis of the Mahalanobis-Taguchi system," *Technometrics*, vol. 45, no. 1, pp. 1-15, 2003.

[10] R. Jugulum, G. Taguchi, S. Taguchi, and J. O. Wilkins, "Discussion: a review and analysis of the Mahalanobis-Taguchi system," *Technometrics*, vol. 45, no. 1, pp. 16-21, 2003.

[11] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.

[12] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the 6th International Symposium on Micro Machine and Human Science*, Nagoya, Japan, 1995, pp. 39-43.

[13] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proceedings of 1997 IEEE International Conference on Systems, Man, and Cybernetics: Computational Cybernetics and Simulation*, Orlando, FL, 1997, pp. 4104-4108.

[14] L. Y. Chuang, C. H. Yang, and J. C. Li, "Chaotic maps based on binary particle swarm optimization for feature selection," *Applied Soft Computing*, vol. 11, no. 1, pp. 239-248, 2011.

[15] X. Xu and Z. Xiao, "Soft set theory oriented forecast combination method for business failure prediction," *Journal of Information Processing Systems*, vol. 12, no. 1, pp. 109-128, 2016.

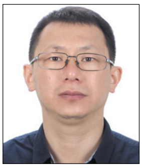
[16] R. Geng, I. Bose, and X. Chen, "Prediction of financial distress: an empirical study of listed Chinese companies using data mining," *European Journal of Operational Research*, vol. 241, no. 1, pp. 236-247, 2015.

Yuping Gu https://orcid.org/0000-0002-1728-1748

She received M.S. and Ph.D. degrees from Nanjing University of Science and Technology, Nanjing, China, in 2004 and 2019, respectively. She is currently a teacher in the School of Management Science and Engineering, Anhui University of Finance and Economics. Her current research interests include pattern recognition and data mining.

Longsheng Cheng https://orcid.org/0000-0003-4727-3711

He received M.S. and Ph.D. degrees from Nanjing University of Science and Technology, Nanjing, China, in 1989 and 1998, respectively. He is currently a professor in the School of Economics and Management, Nanjing University of Science and Technology. His research interests include data mining and Management decision.



Zhipeng Chang https://orcid.org/0000-0003-1157-4133

He is currently a professor in the School of Business, Anhui University of Technology. He received his Ph.D. degree in Nanjing University of Science and Technology, China. His research area includes Mahalanobis-Taguchi System, multiple attribute decision making and pattern recognition.