International Journal of Innovative Computing, Information and Control
Volume **13**, Number **5**, October **2017**

ICIC International ©2017 ISSN 1349-4198
pp. 1759–1767

CLASSIFICATION OF CLASS OVERLAPPING DATASETS BY KERNEL-MTS METHOD

Yuping Gu and Longsheng Cheng

School of Economics and Management
Nanjing University of Science and Technology
No. 200, Xiaolingwei Street, Nanjing 210094, P. R. China
gyp1204@163.com

Received March 2017; revised June 2017

Abstract. *Class overlapping is one of the bottlenecks in data mining and pattern recognition, and affects the classification accuracy and generalization ability directly. In Mahalanobis-Taguchi System (MTS), the normal samples are used to construct reference space, while the abnormal samples are used to verify the validity of the reference space. If there is a class overlapping between the normal samples and the abnormal samples, the result of classification will be affected. In this paper, kernel function and Mahalanobis distance are combined to form the kernel Mahalanobis distance as an improved measurement scale of the MTS. Experimental results show that Kernel-MTS is suitable for class overlapping classification, and it provides better classification accuracy than the conventional methods.*

Keywords: Mahalanobis-Taguchi System (MTS), Kernel function, Class overlapping, Classification

1. **Introduction.**

As the focus in data mining, the problem of classification is getting more and more attention. The methods of classification are widely used in finance, medicine, mechanic and other industries. In these practical problems, due to the objective condition of data collection or the characteristics of data attribute itself, there will be some overlapping areas between the different classes, and the degree of class overlapping directly affects the performance of the classification. Studies have shown that classification errors are usually concentrated in the border areas of different classes, and this is exactly the existence of overlapping regions [1]. There were several studies on class overlapping problems, e.g., Denil and Trappenberg [2] examined the effects of overlap and imbalance on the complexity of the learned model and demonstrated that overlap was a far more serious factor than imbalance in this respect. García et al. [3] studied the overlapping data by artificial generation, but they did not apply the research method in practice. Li and He [4] studied only the problem of overlapping fault diagnosis in two-dimensional space; however, there are many multidimensional data in the practical problem. Xiong et al. [5] studied the cases of concept overlapping, but sample overlapping problems are more common in reality.

Mahalanobis-Taguchi System (MTS) [6] is a highly practical pattern recognition method and was proposed by Dr. Taguchi, who is a well-known Japanese quality engineer scientist. MTS is a diagnosis and forecasting method for multivariate data. MTS regards the signal-to-noise ratio (SNR) and Mahalanobis distance (MD) as the optimization targets, and select the effective variables by using 2-level orthogonal array. In recent years, as a multivariate pattern recognition technology, MTS was widely used in various areas such as industrial production and business management, e.g., Li et al. [7] combined MTS and grey cumulative prospect theory for enterprise information investment and risk decision.

Shakya et al. [8] applied MTS for the online detection of health status of the rolling element bearing. Valarmathi and Palanisamy [9] classified the customers' opinions from the web by using MTS. Although the MTS method has been widely used, it also has disadvantages, such as weak theoretical basis [1], especially when class overlapping exists, the method's classification ability is poor or it cannot be used.

Aiming at the above problem, Kernel Mahalanobis Distance (KMD), which is the Mahalanobis distance in the high dimensional space that is mapped from low dimensional space by the kernel function, is used as a new measurement scale in MTS to solve the problem of classification with overlapping data. KMD is used to map the original samples to feature space (dimensional or infinite-dimensional) implicitly in order to increase the differences between various classes. There have been many studies of kernel function with the original measurement scale [11-14] which is applied in outlier identification, fault diagnosis, quality prediction, etc. And studies have shown that KMD classifier can deal with overlapping and nonlinear problems in the kernel feature space.

The paper is organized as follows. Section 2 starts with preliminaries on kernel function and the kernel Mahalanobis distances. Section 3 introduces Kernel-MTS classification strategies. Section 4 presents experimental comparison on four datasets with some conventional methods. Section 5 gives some conclusions and future directions.

2. **Kernel Mahalanobis Distance.**

2.1. **Review of kernel function.**

The kernel function makes a non-linear mapping from the input space to the feature space, which can make the inner product of the feature space be represented by a function of the input space. Its mathematical expression is shown as $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. The kernel function avoids the curse of dimensionality and reduces the amount of computation greatly. In addition, the dimension of the input space does not affect the kernel function matrix, so the kernel-based methods can deal with the high-dimensional input effectively and the parameters or the form of nonlinear transformation function ϕ need not to be known while only the inner product of low-dimensional space needs to be calculated.

In machine learning theories, the popular kernel functions are shown below.

Gaussian kernel: $k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

Polynomial kernel: $k(x_i, x_j) = \left(a x_i^T x_j + c\right)^d$

Sigmoid kernel: $k(x_i, x_j) = \tanh\left(\alpha x_i^T x_j + c\right)$

where x_i and x_j are the vectors of input space. σ, a, c and d are the customized parameters.

In practice, there is still no general approach to select kernel function. The best kernel function is usually selected by prior knowledge or the cross-validation experiment in which the selection criterion is the minimum training error.

2.2. **Invertible covariance KMD.**

Assume the input space R^P has n sample data: $X = \{x_i\}_{i=1}^n \subset R^p$. The function $\Phi(\cdot)$ is used to map the samples from the input space to the feature space: $\Phi : R^p \to F$, $x \to \phi(x)$. The squared MD between sample vector of input space and reference space with mean μ and covariance matrix C is shown as $d^2(x) = (x - \mu)^T C^{-1}(x - \mu)$. Similarly, MD in feature space can be defined as $d^2(x) = (\phi(x) - \phi_\mu)^T C_\phi^{-1}(\phi(x) - \phi_\mu)$. Feature space is usually high-dimensional or infinite-dimensional, and it is difficult to explicitly express the mapping of sample vector $\phi(x)$, the mean ϕ_μ and the covariance C_ϕ of the mapping of reference space. However, one characteristic of the feature space is that the inner product of any two vectors can be calculated by the kernel function of the corresponding two vectors in the input space. That is, for arbitrary

$x_i, x_j \subset R^p$, there has $\phi(x_i)^T\phi(x_j) = \langle\phi(x_i), \phi(x_j)\rangle = k(x_i, x_j)$. Generally, the kernel matrix of mapped samples can be expressed as $K = \Phi^T\Phi = (k(x_i, x_j))_{n\times n} = (k_{ij})_{n\times n}$.

By choosing kernel functions and its parameters, the mapping from the input space to the feature space is implicitly changed. The following briefly describes the solution steps of invertible covariance KMD [15].

Let the mapped samples in the feature space be represented as $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$, the empirical mean is defined as $\phi_\mu = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) = \frac{1}{n}\Phi\mathbf{1}_n$, where $\mathbf{1}_n$ is a column vector in which the elements are all ones. Then $\tilde{\phi}(x_i) = \phi(x_i) - \phi_\mu$ or more compactly $\tilde{\Phi} = \left[\tilde{\phi}(x_1), \ldots, \tilde{\phi}(x_n)\right] = \Phi - \frac{1}{n}\phi_\mu\mathbf{1}_n^T = \Phi - \frac{1}{n}\Phi\mathbf{1}_n\mathbf{1}_n^T = \Phi H$, where $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, I_n is an identity matrix. So the covariance matrix is $C_\phi = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^T = \frac{1}{n}\Phi HH\Phi^T$, and the centralized kernel matrix is $\tilde{K} = \tilde{\Phi}\tilde{\Phi}^T = HKH$. Usually, the inverse of $\tilde{K}$ does not exist, but its pseudo-inverse matrix $\tilde{K}^-$ can be calculated. In practical application, the calculation of $\tilde{K}^-$ is shown as follows: first, select the kernel function and calculate the kernel matrix, then calculate the $\tilde{K}$, finally, calculate $\tilde{K}^-$ by using the method of singular value decomposition. In this process, the parameter α needs to be set: if the singular value is less than α, $\alpha = 0$; otherwise, $\alpha > 0$. For any x, $x \subset R^p$, the kernel matrix between x and reference space can be expressed as $\mathbf{k}_x = [k(x_1, x), \ldots, k(x_n, x)]^T = \Phi^T\phi(x)$, then $\tilde{\mathbf{k}}_x = \tilde{\Phi}^T\tilde{\phi}(x) = H\left(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n\right)$, so the KMD can be calculated as follows:

$$KMD = d_{IC}^2(x) = \tilde{\phi}(x)^T C_\phi^{-1}\tilde{\phi}(x) = n\tilde{\mathbf{k}}_x^T\left(\tilde{K}^-\right)^2\tilde{\mathbf{k}}_x$$

3. Proposed Methodology.

In traditional MTS, the validity of the reference space is verified by abnormal samples. When the difference between normal and abnormal MDs is small, the classification effect will be unsatisfied. If this situation occurs, there may have class overlapping between normal and abnormal samples. KMD should be used to form a new measurement scale instead of MD and thus Kernel-MTS would be used for data classification. Similar to the traditional MTS, the application of Kernel-MTS also can be carried out in four stages, shown as follows.

Stage 1: construct the reference space. Define the variables under healthy or normal conditions. Collect the normal samples and normalize them. Select the basic kernel function or construct compound kernel function, and use the formula in Section 2.2 to calculate the KMD for all samples in the normal group.

Stage 2: confirm the validity of the reference space. Identify the unhealthy or abnormal condition, and normalize them by using the mean and standard deviation of normal group, and calculate the KMD of abnormal samples. If the reference space is valid, the KMD of abnormal samples will be larger than normal samples. Based on this, the validity of the reference space can be judged.

Stage 3: identify valid variables. The valid variables are selected by using orthogonal array and larger-the-better SNR. Each combination of variables in the orthogonal array yields an SNR which is calculated by the KMDs of abnormal samples. Depending on the difference in SNR at different variable levels, the set of valid variables can be identified.

Stage 4: use valid variables to diagnose. According to the reference space composed of valid variables, the KMDs of unknown samples are calculated. By the comparison between the calculated KMD value and threshold value, the classification, diagnosis and prediction can be carried out. In the application of Kernel-MTS, the determination of threshold is very important because it can directly affect the classification effect. There are several ways to determine the threshold in traditional MTS. Taguchi and Jugulum

[6] used quadratic loss function to determine the threshold but because of the large uncertainty of quantitative loss, the method is difficult to implement in practice. Su and Hsiao [16] proposed a probabilistic threshold model with the lowest error rate of normal samples; however, this model has complicated computational process and exits pending parameters. In practice, exhaustive method is usually used to determine the optimal threshold, in order to obtain a higher overall classification accuracy. However, the exhaustive method takes a long time, and it may reduce the recoquition accuracy of normal or abnormal samples in order to achieve higher overall accuracy.

In this paper, the f-max method is proposed to determine the threshold. Assume T is the threshold, the KMD of normal samples is defined as KMD_i $(i = 1, 2, \ldots, n)$, and the KMD of abnormal samples is defined as KMD_j $(j = n + 1, n + 2, \ldots, n + m)$.

Thus,

$$n_{error} = \sum_{i=1}^{n} k_i, \quad k_i = \begin{cases} 1, & \text{if } KMD_i > T \\ 0, & \text{if } KMD_i \leq T \end{cases} \qquad i = 1, 2, \ldots, n$$

$$m_{error} = \sum_{j=n+1}^{n+m} k_j, \quad k_j = \begin{cases} 1, & \text{if } KMD_j < T \\ 0, & \text{if } KMD_j \geq T \end{cases} \qquad j = n + 1, n + 2, \ldots, n + m$$

Then the classification accuracy of normal samples is $f_1 = (n - n_{error})/n \times 100\%$, and the classification accuracy of abnormal samples is $f_2 = (m - m_{error})/m \times 100\%$. Let $f = f_1 \times f_2$, and make f maximization by adjusting the value of T. When the f value is the maximum value, the corresponding T value can be selected as the threshold. The f-max method is also suitable for MTS.

The flow chart of Kernel-MTS method is shown as Figure 1.

4. Experimental Comparison.

4.1. Datasets and experimental methods.

Four common classification datasets in UCI database are chosen and their basic information is shown in Table 1.

Table 1. Datasets information

Datasets name	No. of variables	No. of samples	Normal sample representation/No.	Abnormal sample representation/No.
Statlog (Heart)	13	270	absence/150	presence/120
Ionosphere	34	351	Good/225	Bad/126
Glass identification	9	146	Float_processed/70	Non_float_processed/76
Forest type mapping	9	354	Sugi forest/195	Mixed deciduous forest/159

The experiment uses 5-fold cross-validation method, that is, each dataset is randomly split into 5 mutually exclusive subsets of approximately equal size, four subsets are selected as training sets of each experiment, and the remaining one subset is as a test set. Traditional MTS, decision tree C4.5 algorithm, Support Vector Machine (SVM), Naive-Bayes (NB), k-Nearest Neighbor (k-NN) and Kernel-MTS are used simultaneously for comparative study. The parameter σ of Gaussian kernel in Kernel-MTS is selected when the average test error is smallest in the test set. The parameter α is set as 0.5 to solve the pseudo inverse matrix of the centralized kernel matrix $\tilde{K}$. The f-max method is used to determine the threshold in both MTS and Kernel-MTS algorithm. The parameter k is set as 5 in the k-NN algorithm. The parameters in C4.5, SVM, NB and other algorithm are determined by multiple cross-validations on several subsets.

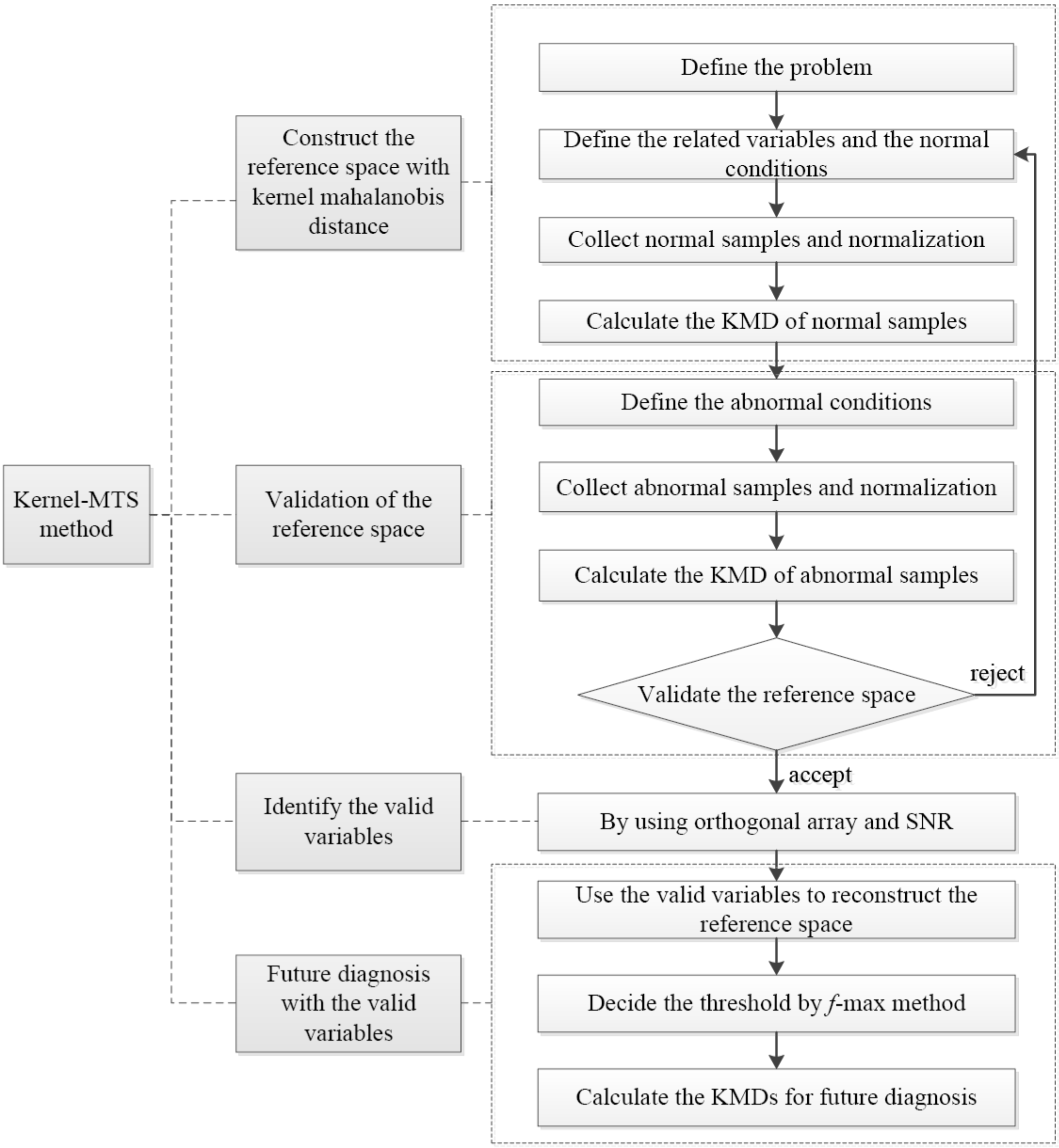
Kernel-MTS method
Construct the reference space with kernel mahalanobis distance
Validation of the reference space
Identify the valid variables
Future diagnosis with the valid variables
Define the problem
Define the related variables and the normal conditions
Collect normal samples and normalization
Calculate the KMD of normal samples
Define the abnormal conditions
Collect abnormal samples and normalization
Calculate the KMD of abnormal samples
Validate the reference space
reject
accept
By using orthogonal array and SNR
Use the valid variables to reconstruct the reference space
Decide the threshold by f-max method
Calculate the KMDs for future diagnosis

Figure 1. The flow chart of Kernel-MTS method

The calculation of the above methods is carried out by R software. Evaluating the classification effect of each method on each dataset is based on the mean of the results of the 5-fold cross-validation experiments. The evaluation metrics are the classification accuracy of the normal samples, the abnormal samples and the overall samples and F-measure.

A two-class problem confusion matrix is shown as Table 2. In the table, True represents normal sample, False represents abnormal sample. TP represents the number of normal samples which is predicted correctly, TN represents the number of abnormal samples which is predicted correctly, FP represents the number of normal samples which is predicted incorrectly, and FN represents the number of abnormal samples which is predicted incorrectly.

The commonly used evaluation metrics are shown as follows:

(1) Sensitivity (TPR) = TP/(TP+FN), which is the classification accuracy of normal samples. (2) Specificity (TNR) = TN/(TN+FP), which is the classification accuracy of abnormal samples. (3) Accuracy (Acc) = (TP+TN)/(TP+TN+FP+FN), which is the

Table 2. Confusion matrix

	Predicted Positive	Predicted Negative
Actual True	True Positive (TP)	True Negative (TN)
Actual False	False Positive (FP)	False Negative (FN)

classification accuracy of overall samples. (4) F-measure = 2×TP/(2×TP+FP+FN). The higher the four metrics above, the better the classification effects.

4.2. **Results analysis.** The experimental process and the results are descripted by using the Statlog dataset as an example, and the remaining datasets only show the final results. In the 5-fold cross-validation experiment of Statlog dataset, the number of training set is 216, in which the number of normal sample is 120 and the number of abnormal sample is 96. The number of test set is 54, in which there are 30 normal samples and 24 abnormal samples. MD and KMD density distribution curve of normal samples and abnormal samples by using MTS and Kernel-MTS under the valid variables are shown as Figures 2 and 3. In these figures, 0 represents normal samples and 1 represents abnormal samples.

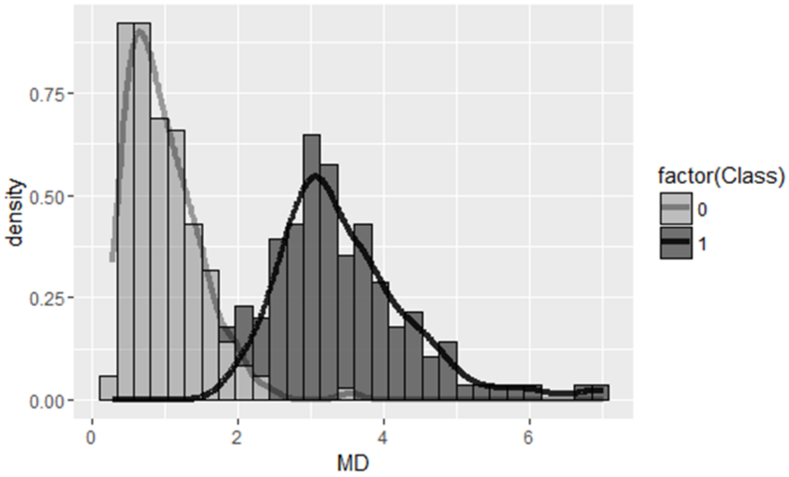

Figure 2. The density curve in MTS

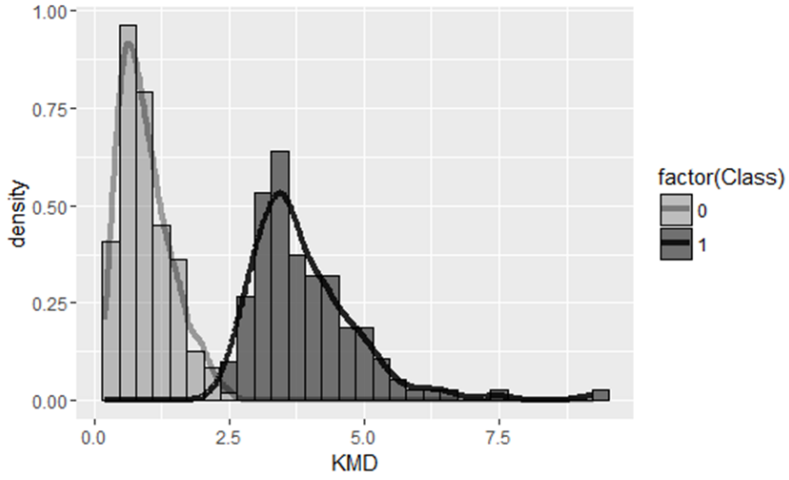

Figure 3. The density curve in Kernel-MTS

As shown in the figures, there exist class overlapping in Statlog dataset, which also can be verified by the low classification accuracy of the conventional classification algorithm. The classification accuracy of each algorithm on the Statlog dataset is shown in Table 3. The classification effect of Kernel-MTS algorithm is better than the other methods in which *Acc* metric is particularly prominent. The classification effect of MTS is at a moderate level. By using 5-fold cross-validation, the average number of deleted variables in MTS and Kernel-MTS are 3.8 and 5.2 respectively. So the Kernel-MTS algorithm has a better result in dimension reduction.

Table 3. Classification effect of Statlog dataset

	TPR (%)	***TNR*** (%)	***Acc*** (%)	***F-measure*** (%)
Kernel-MTS	90.87	82.70	87.58	86.49
MTS	84.26	78.88	80.53	81.78
SVM	86.55	78.45	82.57	84.21
***k*-NN**	75.33	60.63	68.15	72.29
C4.5	83.98	71.67	78.52	81.14
NB	86.15	81.67	83.54	84.68

For the datasets of forest type mapping, glass identification and ionosphere, the comparisons of classification effect are shown in Tables 4 to 6.

Table 4. Classification effect of forest type mapping dataset

	TPR (%)	***TNR*** (%)	***Acc*** (%)	***F-measure*** (%)
Kernel-MTS	95.59	89.52	92.95	93.86
MTS	94.85	87.62	91.70	92.80
SVM	95.57	86.84	91.77	92.51
***k*-NN**	95.57	79.45	88.55	89.95
C4.5	93.43	78.52	86.93	88.55
NB	89.15	84.99	87.34	88.50

Table 5. Classification effect of glass identification dataset

	TPR (%)	***TNR*** (%)	***Acc*** (%)	***F-measure*** (%)
Kernel-MTS	92.03	88.95	90.14	90.31
MTS	87.65	84.21	85.19	85.03
SVM	88.12	84.33	85.93	85.32
***k*-NN**	92.71	76.91	84.38	84.91
C4.5	80.15	85.55	82.84	81.08
NB	89.30	72.31	81.32	82.41

4.3. **Summary.** Through the study of the above four examples, except the ionosphere dataset, Kernel-MTS has the best results with the other three datasets, especially with more serious overlapping datasets, which shows Kernel-MTS is a reasonable and effective method under the situation of class overlapping. The traditional MTS has a good classification effect on the datasets without class overlapping, while it is not good with class overlapping. Other methods have different performance on different datasets. Compared with the other datasets, the ionosphere dataset is imbalanced which may be the reason why the classification of Kernel-MTS is not optimal.

Table 6. Classification effect of ionosphere dataset

	TPR (%)	***TNR*** (%)	***Acc*** (%)	***F-measure*** (%)
Kernel-MTS	93.63	86.67	90.90	91.63
MTS	92.97	86.53	87.64	88.34
SVM	96.78	87.09	92.3	93.67
***k*-NN**	96.78	60.43	83.34	88.01
C4.5	92.44	80.12	88.03	90.83
NB	79.11	86.49	81.76	84.73

In addition to classify the data, MTS and Kernel-MTS also can be used to identify valid variables. In general, Kernel-MTS is better than MTS and the reason is that Kernel-MTS uses the kernel function to map the samples to the high-dimensional space, and thus eliminates the possible interference variables. So a better classification effect can be achieved with fewer valid variables by Kernel-MTS.

5. Conclusions and Future Directions.

In this paper, based on the traditional MTS, the kernel function and the mahalanobis distance are combined as a new measurement to deal with the class overlapping. Through the study of four datasets, the results show that the classification effect and the dimension reduction of Kernel-MTS are superior to the traditional MTS, which not only solves the classification of overlapping class, but also expands the application of MTS, so it has big reference value to the actual classification problem.

In this paper, the future directions are:

(1) Kernel-MTS improved the classification effect of three balanced datasets while the classification effect of an imbalanced dataset is not the best. How to deal with the datasets simultaneously existing class imbalance should be considered at the same time.

(2) In the process of solving KMD, the parameters of α and σ are chosen by experience. Particle swarm optimization, genetic algorithm and other optimization methods can be considered to optimize the parameters.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China (No. 71271114). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] S. Alshomrani, A. Bawakid, S. O. Shim, A. Fernández and F. Herrera, A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets, *Knowledge-Based Systems*, no.73, pp.1-17, 2015.

[2] M. Denil and T. Trappenberg, Overlap versus imbalance, *Proc. of the 23rd Canadian Conference on Artificial Intelligence*, pp.220-231, 2010.

[3] V. García, R. A. Mollineda and J. S. Sánchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, *Pattern Analysis and Applications*, vol.11, nos.3-4, pp.269-280, 2008.

[4] T. E. Li and Z. He, Identification of overlapping faults based on PCA-shaping and LDA, *Journal of Systems Engineering*, vol.27, no.5, pp.712-718, 2012.

[5] H. T. Xiong, J. J. Wu, H. P. Liu and L. Liu, Towards classification with class overlapping, *Journal of Management Sciences in China*, vol.16, no.4, pp.8-21, 2013.

[6] G. Taguchi and R. Jugulum, *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*, John Wiley & Sons, 2002.

[7] C. B. Li, J. H. Yuan and P. Gao, Risk decision-making based on Mahalanobis-Taguchi system and grey cumulative prospect theory for enterprise information investment, *Intelligent Decision Technologies*, vol.10, no.1, pp.49-58, 2016.

[8] P. Shakya, M. S. Kulkarni and A. K. Darpe, A novel methodology for online detection of bearing health status for naturally progressing defect, *Journal of Sound and Vibration*, vol.333, no.21, pp.5614-5629, 2014.

[9] 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.

[10] W. H. Woodall, R. Koudelik, K. L. Tsui et al., A review and analysis of the Mahalanobis-Taguchi system, *Technometrics*, vol.45, no.1, pp.1-15, 2003.

[11] Y. Li, H. Zhang and X. Ji, Triangular Hermite kernel extreme learning machine, *International Journal of Innovative Computing, Information and Control*, vol.12, no.6, pp.1893-1904, 2016.

[12] Y. L. Chen, L. J. Lu and X. B. Li, Kernel Mahalanobis distance for multivariate geochemical anomaly recognition, *Journal of Jilin University (Earth Science Edition)*, no.1, pp.396-408, 2014.

[13] M. R. Vazifeh, P. Hao and F. Abbasi, Fault diagnosis based on multikernel classification and information fusion decision, *Computer Technology and Application*, vol.4, no.8, 2013.

[14] Q. Li, Q. Du, W. Ba and C. Shao, Multiple-input multiple-output soft sensors based on KPCA and MKLS-SVM for quality prediction in atmospheric distillation column, *International Journal of Innovative Computing, Information and Control*, vol.8, no.12, pp.8215-8230, 2012.

[15] B. Haasdonk and E. Pźkalska, *Classification with Kernel Mahalanobis Distance Classifiers, Advances in Data Analysis*, Springer Berlin Heidelberg, 2009.

[16] C. T. Su and Y. H. Hsiao, An evaluation of the robustness of MTS for imbalanced data, *IEEE Trans. Knowledge & Data Engineering*, no.10, pp.1321-1332, 2007.