International Journal of Grid and Distributed Computing
Vol.7, No.1 (2014), pp.67-76
http://dx.doi.org/10.14257/ijgdc.2014.7.1.07

Performance Analysis and Coding Strategy of ECOC SVMs

Zhigang Yan[1, 2] and Yuanxuan Yang [1,2]

[1] *School of Environmental Science and Spatial Informatics, China University of Mining and Technology, Xuzhou, Jiangsu, P.R.China*
[2] *Jiangsu Key Laboratory of Resources & Environmental Information Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, P.R.China*
Corresponding author: Zhigang Yan, zhg-yan@163.com

Abstract

The theoretical upper bound of generalization error for ECOC SVMs is derived based on Fat-Shattering dimensionality and covering number. The factors affecting the generalization performance of ECOC SVMs are analyzed. From the analysis, it is believed that in real classification tasks, the performance of ECOC depends on the performance of the classifiers corresponding to its coding columns, which is irrelevant to the mathematical characteristics of the ECOC itself. The essence of ECOC SVMs is how to construct an optimal voting machine consisting of a number of SVMs, how to choose Sub-SVMs which have better generalization ability, and how to determine the number of Sub-SVMs taking part in voting, that is the most important issue. Data sets including "Segment" are selected for test. All the ECOC code columns are constructed using an exhaustive technique. A Sub-SVM is trained for each code column, and the generalization ability of each Sub-SVM is evaluated by classification intervals and error rates estimated by cross validation. Then, all the ECOC code columns are sorted by the generalization performance of Sub-SVMs. Three categories of ECOC SVMs, including superior, inferior and ordinary categories, are constructed from the sorted ECOC code columns, by using forward, backward and original sequences. Experimental results show that the performance of ECOC SVMs which consist of Sub-SVMs with better generalization ability is better and vice versa, which validates our view and points out the direction for improving ECOC SVMs.

Keywords: *ECOC, SVM, Generalization Ability, Code Matrix*

1. Introduction

Numerous supervised learning algorithms are designed for two-class problems, for example, support vector machines (SVM) [1]. However, in real applications, many problems are multiclass problems. Therefore, generalizing SVM to deal with multiclass problems is still one of important research activities in machine learning. Currently, the usual practice is to convert a multiclass problem into a number of two-class problems and then combine them in some way to realize classification into multiple classes. Error Correcting Output Codes (ECOC) is one of the commonly used combination way [2], which is called ECOC SVMs. However, there is not a general coding method which can generate appropriate ECOC for any class number. Furthermore, the existing coding strategy is based on the research on mathematical features of code matrix, which ignores the fundamentals of classification, making it difficult to progress for ECOC SVMs and their applied research.

In this study, it is believed that, in real classification problems, different coding sequences of ECOC SVMs have different meanings. The performance of coding does not depend on

ISSN: 2005-4262 IJGDC

code matrix itself; instead, it depends on the performance of the real classification problems corresponding to the code columns. According to this viewpoint, we attempt to investigate ECOC SVMs from real classification problems in this study, which points out the direction for improving ECOC SVMs.

2. Code Matrix of ECOC and its Corresponding Classifiers

2.1 Code Matrix of ECOC

ECOC is a coding matrix consisting of {0,1} shown in Table 1, denoted as MQ×S. In multiclass problems, row Q represents the class number of samples, while column S represents the number of classifiers to be trained. When Mqs=1(Mqs=0), this sample is positive (negative) for the q-th class and the s-th classifier fs. The working process of ECOC is divided into two phases: training and classification. In the training phase, the classifier f(x)=(f1(x),…,fs(x)) is trained according to the above-mentioned principle; while in the classification phase, for a new sample X, the distances between output vectors and the class vectors are calculated. Then class with the minimum distance is the classification result, which is given by:

$$K = \arg\min_{q\in[1..Q]}(d(M_q, f(X)) \quad (1)$$

where K is the class of X, and d is the distance function. The Hamming Distance (HD) is usually used:

$$d(M_q, f(x)) = \sum_{s=1}^{S} \frac{\left|2m_{qs} - \mathrm{sgn}(f_s) - 1\right|}{2} \quad (2)$$

Table 1. All Possible ECOC Columns for a 4-Class Problem

Class	Code Word						
	f_1	f_2	f_3	f_4	f_5	f_6	f_7
C_0	0	0	0	0	0	0	0
C_1	1	1	1	1	0	0	0
C_2	1	1	0	0	1	1	0
C_3	1	0	1	0	1	0	1

For ECOC, when the coded rows are the same, the classes corresponding to the rows cannot be identified; when the coded columns are the same, they correspond to the same classifier, therefore deleting a column does not affect the output; when the code of two columns are complementary, the outputs of their corresponding classifiers are complementary, therefore they are identical; columns of all “0” or all “1” are make no sense, because they cannot be used to train classifiers. In one word, an available ECOC must satisfy the following conditions:

(1)The rows of the coding matrix are not correlated, and neither correlated nor complementary are the columns of the coding matrix;

(2)None of the columns is all “0” or all “1”.

(3)For a k-class problem, the coding length L must satisfy $\log_2^k < L \le 2^{k-1} - 1$;

According to the coding theory, for an error correction code with minimum HD d, $[(d-1)/2]$ bits of error can be corrected. Therefore, for an output code with error correction ability, the HD between code words should be larger than 3.

Dietterich proposed four commonly used ECOC coding methods [2], including Exhaustive Codes, Column Selection from Exhaustive Codes, Randomized Hill Climbing Codes and BCH Codes. In addition, Crammer and Singer proposed the concept of continuous coding [3]. Utschick proposed expectation maximization coding algorithm [4], in which the ECOC is selected by constructing maximized objective function. Ludmila and Kuncheva used hybridization and mutation in evolution algorithm to derive new ECOC codes from random ones [5]. The recent research about ECOC is a general coding method - searching coding method which was proposed in reference [6]. The method is not only suitable for problems of any class number, but also can automatically generate alternative codes according to different criteria, including class numbers and minimum HDs. However, it cannot deal with the problem caused by identical columns.

For the evaluation of coding performance, Francesco believes that the performance of ECOC is related to many factors, including: the similarity of coding words, the performance of the classifiers, the complexity of the real problems, the choice of classifiers, and the correlation of the coding columns, *etc.*, [7]; Xia believes the performance of ECOC is related to coding length, the minimum HD between code words, and the distribution order of the code words [8].

It can be seen from the above that the evaluation of ECOC coding performance and the application in classification start from coding itself, while attention is seldom paid on classification. Next, we introduce the ECOC SVMs in real classification problems.

2.2. SVM

SVM is a machine learning method based on statistical learning theory. To resolve the pattern recognition problem, a calculable recognition function $y = f(x), x \in R^n, y \in \{-1,1\}$ is found. For the given k samples $(x_1, y_1),(x_2, y_2),\cdots (x_k, y_k), x \in R^n, y \in \{-1,1\}$, a hyperplane (decision surface) needs to be found, namely, $Wx + b = 0, W \in R^n, b \in R$, and the corresponding recognition function is:

$$f(x) = sign((Wx) + b) \tag{3}$$

The decision surface should meet the following constraints:

$$y_i[Wx_i + b] \geq 1 - \xi_i, i = 1, 2, \cdots, k \tag{4}$$

The optimal decision surface should meet the requirement that the smallest distance from the two classes of samples to the decision surface is the biggest, hence, the classification problem becomes that the condition of $\xi_i \geq 0$ should be met and the minimum problem of Formula (4), namely:

$$\min : \tau(W) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{k} \xi_i \tag{5}$$

The first item in the formula makes the smallest distance from the two classes of samples to the decision surface the biggest, while the second item makes the error the minimum, and the constant C splits the difference of two above. This optimization problem with constraints

can be resolved with Lagrangian approach, and the corresponding classification function can be changed to:

$$f(x) = sign(\sum_{i=1}^{k} \alpha_i y_i (x_i \bullet x) + b) \tag{6}$$

For the non-linear separable condition, a non-linear function ϕ can be found, and then the data are mapped to a high dimensional feature space, in which an optimal hyperplane is established, and the corresponding classification function is as follows:

$$f(x) = sign(\sum_{i=1}^{k} \alpha_i y_i (\phi(x) \bullet \phi(x_i)) + b) \tag{7}$$

Only point multiplication algorithm $K(x, y) = \phi(x) \bullet \phi(y)$ in the high dimensional feature space is considered in SVM theory, in which $K(x, y)$ is called kernel function, and the function ϕ is not used directly, hence, formula (7) can be transferred into formula (8):

$$f(x) = sign(\sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b) \tag{8}$$

The common kernel function includes: linear kernel function, $K(x, y) = (x \bullet y)$; RBF kernel function, $k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$.

2.3. ECOC SVMs

Combining SVMs with ECOC to classify multiple classes, we have ECOC SVMs. The upper bound of the generalization error for ECOC SVMs is derived in reference [8] based on the concept of Fat-Shattering dimension and covering number. Assuming m samples can be correctly classified by k-class ECOC SVMs, with ECOC coding length being L, minimum HD between code words being d, and the sorted SVM classification intervals in descendent order denoted by $\gamma_1, \gamma_2, \cdots, \gamma_L$, the generalization error ECOC SVMs, with probability at least $1 - \delta$, is no larger than:

$$\frac{130R^2}{m}\left(D' \log_2(4em)\log_2(16m) + \log_2 \frac{2(2m)^M MNK!}{\delta}\right) \tag{9}$$

where $D' = \sum_{i=1}^{L} \frac{1}{\gamma_i^2}$, R is the minimum radius of enclosure ball, $M = [L - (d-1)/2]$, N is the number of codes with coding length L and HD d between codes. Each group has K code words. It is believed in [8] that:

(1) Given a fixed code length, the longer minimum HD between codes, the better generalization ability of ECOC SVMs;

(2) Given a fixed minimum HD between codes, the longer code length, the worse generalization ability of ECOC SVMs;

(3) Once the code length and the minimum HD between codes are fixed, there exists optimal allocation order for code words which guarantees the ECOC SVMs the best generalization ability.

The relation between the generalization ability of ECOC SVMs and the minimum HD between codes and the code length is discussed in [8]. However, no discussion on the relation between the minimum HD between codes and the code length is given. Furthermore, it is believed in [8] that there do exist optimal coding sequence but no solution is provided according to coding itself. Clearly, it is difficult to find a way to determine the code length and the code sequences from a mathematic point of view. Exhaustive search is definitely unfeasible and it does not provide a reasonable explanation to coding. Intuitively, the answer should be found in the classification problems themselves.

3. New Understandings about ECOC SVMs

Analyzing formula (9) again, one knows that D, M and N affect the upper bound of the generalization error of ECOC SVMs. In formula (9), $M = \lfloor L - (d-1)/2 \rfloor$, the minimum HD d is related to code length L. Generally, the larger L, the larger d. However the increment of $(d-1)/2$ is less than or equal to that of L. Therefore, M is non-decreasing; N is also related to L. It increases when L increases. $D = \sum_{i=1}^{L} \frac{1}{\gamma_i^2}$, with the increase of code length L, D is increasing. Therefore, the generalization ability of ECOC SVMs decreases when L increases. In this study, we believe that ECOC SVMs should have enough Sub-SVMs for decision. That is to say, L should be big enough. However, a bigger L may harm the performance of ECOC SVMs. Thus, the value of L should be a compromise. When L is determined, the effect of D on the generalization ability of ECOC SVMs is major. Select L Sub-SVMs with good generalization ability, then the generalization ability should be good if one constructs ECOC SVMs with these Sub-SVMs. If it is impossible to evaluate the generalization ability of each Sub-SVM, or the difference between each is insignificant, the effect of M and N on the performance of ECOC SVMs, which is the conclusion of reference [8].

A "voting" process is used to vividly describe the above analysis. The essence of ECOC SVMs is to train a number of two-class SVMs, then determine the class of an unknown sample according to the classification results of these two-class SVMs. Using minimum HD to determine the class of a sample is equivalent to voting. In the voting stage, each Sub-SVM votes for a number of classes which it supports; in the calling stage, the sample class is the class which most corresponds to the results of Sub-SVMs. Each column of the codes corresponds to a Sub-SVM, therefore, the process of constructing ECOC determines which of the Sub-SVMs have the voting right. Clearly, it is important to give voting right to those Sub-SVMs with good generalization ability. Thus, the coding problem is actually how to construct an optimal voting machine consisting of a number of two-class SVMs, where how to select Sub-SVMs with good generalization ability and how to determine the number of Sub-SVMs taking part in voting are two important factors. Next, experiments are used to validate the viewpoint.

It is pointed out in [1] that VC dimension and the classification interval γ when linearly separable are the criteria for the generalization ability of an SVM. However, it is difficult to determine the VC dimension. Therefore VC dimension is difficult to deal with and apply. The classification interval γ is a reliable criterion for evaluating the generalization ability of an SVM, but it needs the precondition that the SVM can linearly separate samples, which is usually difficult to satisfy. When samples are not linearly separable but are linearly separable after being mapped into a high dimensional space, the generalization ability of an SVM is described by the classification interval in high dimensional space. In this situation, the classification interval of an SVM is $\gamma = 2/\|W\|$, where W is the normal vector of the

classification hyperplane of the SVM, which is calculable. However it is impossible to directly calculate W. Noticing the duality of (6) and (8), one can obtain:

$$\|W\|^2 = 2\sum_{i=1}^{k} \alpha_i + 2C\sum_{i=1}^{k} \xi_i - \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (10)$$

When samples are completely separable, $2C\sum_{i=1}^{k} \xi_i = 0$, then (10) is simplified as:

$$\|W\|^2 = 2\sum_{i=1}^{k} \alpha_i - \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (11)$$

$\|W\|^2$ can be calculated according to (11), thereby the classification interval γ in high dimensional space can also be obtained. When samples are partially separable, we use the error rate E_i of the cross validation to evaluate the generalization ability of SVMs. The smaller E_i, the better generalization ability. To reduce the misclassification, the classification interval γ_i is also considered when using E_i as a evaluation criterion. But the classification interval now includes the misclassified samples. The effect of misclassified samples should be eliminated when calculating.

Data sets from UCI database are selected for test, including Segment, Landsat, Optdigits, Zoo, Page Blocks, *etc.*, Linear kernel and RBF kernel are used in the test. The process of the test is as follow: Given k classes of samples, all the ECOC code columns are constructed by exhaustive method, totaling $2^{k-1}-1$ columns. For each column of code, train the Sub-SVMs. Then sort the code columns by the generalization ability of Sub-SVMs, according to the following rules:

(1) When samples are linearly separable, the linear kernels are chosen. They have lower VC dimensions and better generalization ability compared with RBF kernels;

(2) When samples are linearly inseparable but separable if RBF kernels are used, the RBF kernels are chosen;

(3) When samples are completely inseparable, kernel functions with higher accuracies are chosen;

(4) When samples are separable, sort the classification interval γ_i by descending order;

(5) When samples are inseparable, sort the error rate E_i by ascending order, meantime γ_i are referred.

A number of code columns in positive sequence are chosen from the sorted ECOC code columns to train the ECOC SVMs, then the same number of code columns in reverse order are chosen as comparison group. The original order of the exhaustive code is kept unchanged. The same number of code columns are successively chosen as reference groups. The experimental results of the selected data sets are basically the same. Taking the Segment data set as an example, the results are shown in Figure 1. There are 7 classes in Segment data set, with each sample having 19 features. Each class provides 30 training samples and 300 test samples. The code length ranges from 3 to 63. In the experiments, ECOC SVMs consisting of 10 to 63 code columns are tested. RBF kernels have higher accuracy, so they are chosen in the experiments. To facilitate the test, same parameters are used for all Sub-SVMs.

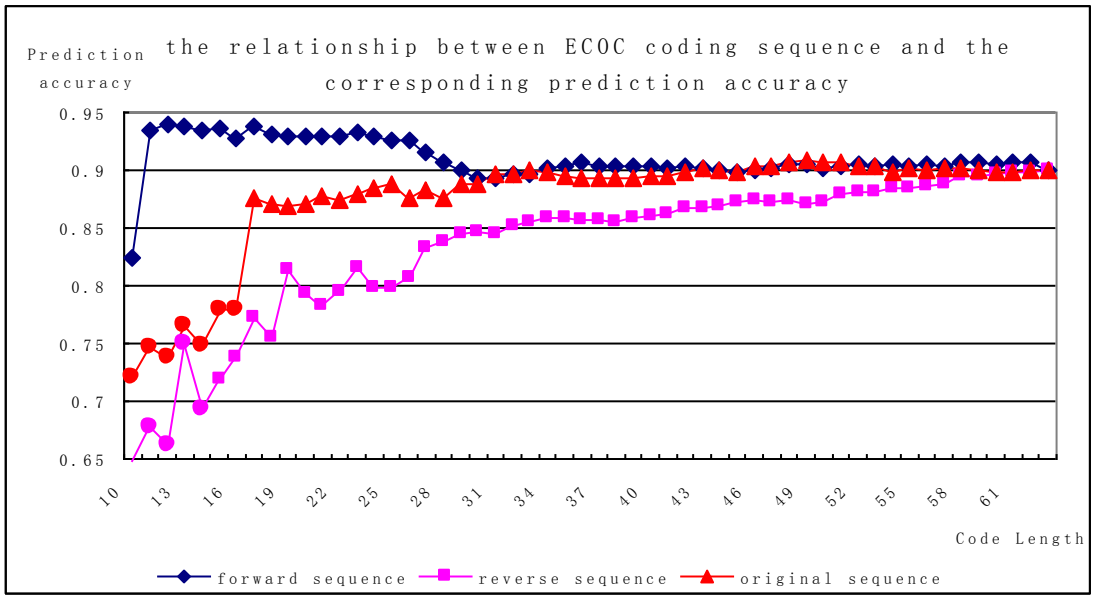

Figure 1. Relationship between ECOC SVMs Encoding Method and their Prediction Accuracy for "Segment" Data Set

It can be seen in Figure 1 that the prediction accuracy of ECOC forward sequences is much higher than that of reverse sequences, while the original sequences have the medium accuracy. The original sequences can be viewed as the prediction of random ECOC, while the reverse sequences have the worst prediction results. The forward sequences have the best prediction results. In addition, the more overlapping code between forward and reverse sequences, the closer prediction accuracies they have. When code length increases to a certain degree, the prediction accuracy of forward sequences decreases and becomes stable, while the accuracy of original sequences increases with fluctuations and finally becomes stable. However the accuracy of reverse sequences keeps increasing. The results suggest that when code length increases, if the generalization ability of the SVMs corresponding to the newly added columns are strong, the coding performance improves, like reverse sequences; Conversely, if the generalization ability of the SVMs is weak, the coding performance degrades, like forward sequences; in original sequences, code lengths are short at the beginning, which makes the coding performance bad, however with the increase of code lengths, the minimum HD between codes increases, improving the generalization ability. But if the code lengths still increase, more code columns with weak generalization ability exist, which stops the overall performance from increasing. It is believed in [8] that the generalization ability of ECOC SVMs depends on the first $[L-(d-1)/2]$ high performance SVMs, irrelative to the rest (d-1)/2 SVMs. However we further show that when generalization ability is good, insufficient coding numbers will also affect the generalization ability of ECOC SVMs. In this situation, the generalization ability of ECOC SVMs is relative to the Sub-SVMs with bad generalization ability. More importantly, a coding strategy is derived in this study showing how to construct ECOC with good generalization ability.

It can be seen in Figure 2 that, the performance of ECOC SVMs neither necessarily increases with the increase of the minimum HD, nor with the increase of code length. Instead, it has a complex relationship with both of them.

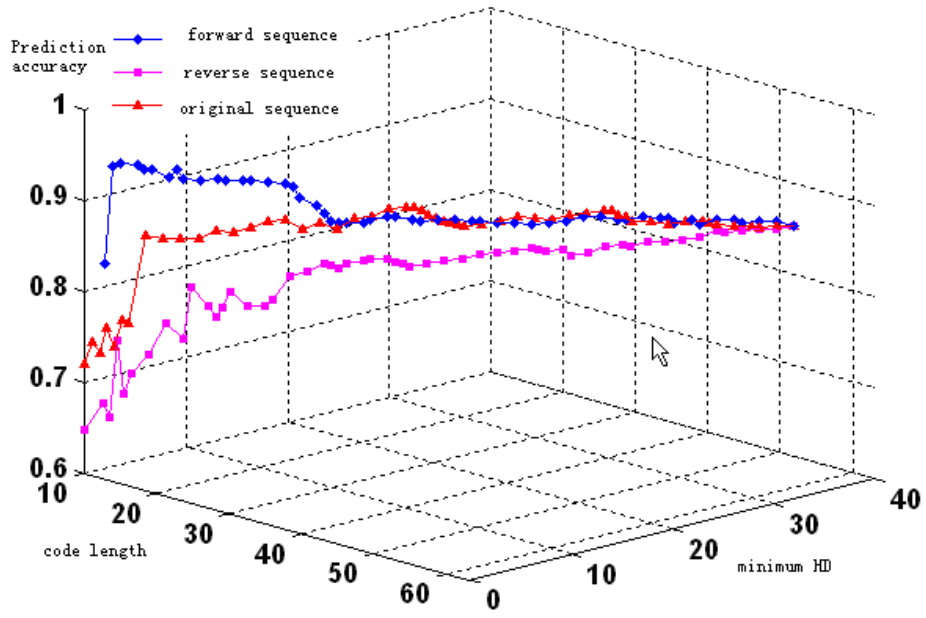

Figure 2. Relationship between Prediction Accuracy, Code Length, and the Minimum Hamming Distance

The viewpoint in this study is validated through real tests. That is, the generalization ability of SVMs corresponding to code columns has the most significant effect on ECOC performance, while code lengths and minimum HD between codes are both mathematical features represented by coding. The key factor in improving the performance of ECOC SVMs is to find Sub-SVMs with good generalization ability. When impossible to find such Sub-SVMs or the generalization abilities of all Sub-SVMs are the same, the performance of ECOC SVMs can be considered from the coding point of view. Then code length, minimum HD between codes, allocation order of codes, correlations between code columns can be the evaluation criteria of the generalization ability of ECOC SVMs, which is discussed in [8].

The exhaustive coding method can assure the code columns are neither correlated nor complementary. However when code length is short, same columns may exist in ECOC, make it impossible to determine the class of a sample. In this situation, remedial measures should be taken, that is, training additional SVM classifiers corresponding to the identical code words, in order to judge the decision results of ECOC SVMs. In Figure 2, there are cases in which the minimum HD between codes is 0 when the reverse sequence and the original sequence are both short, implying that there are identical codes. Strictly speaking, the ECOC is wrong in these cases. However for convenience, this part of code is reserved after taking measures to deal with it, marked by "●" in Figure 1.

4. Conclusions and Discussions

Starting from the essence of problems, ECOC SVMs is analyzed in this study. New constructing method is proposed. The main conclusion and problem are as below:

1. The performance of ECOC SVMs depends on the performance of its corresponding Sub-SVMs, while the mathematical features represented by coding are secondary. When impossible to evaluate the performances of Sub-SVMs or the performances are the same, the code lengths, the minimum HD between codes, the allocation order of code words, and the

correlations between codes can be the criteria of ECOC. For the error correction ability, it is believed in this study that for optimal classifiers, less error is primary, while error correction is secondary.

2. What needs to further solve is that: (1) currently, there are still no exact theory about the evaluation of the SVM performance because kernel functions, parameters and sample spaces are different. Also, there is still no exact theory about the comparability of the generalization ability of each Sub-SVM; (2) What is the appropriate code length of ECOC SVMs? What is the relation between the code length of ECOC SVMs and the generalization ability of each Sub-SVM? And how to conveniently and rapidly construct a reasonable code matrix? These problems are still to be investigated. From the initial results, code can be short for Sub-SVMs with better generalization abilities, or quality is more important than quantity; when the performance of Sub-SVMs is impossible to evaluate, code length should be longer, or compensate quality by increasing quantity. But the code length should be appropriate, by no means the longer the better.

Acknowledgements

This work was supported by a grant from Natural Scientific Fund of China (No. 41271445) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] V. N. Vapnik, "The Nature of Statistical Learning Theory", Springer, New York, USA, **(1995)**.
[2] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes", Journal of Artificial Intelligence Research, vol. 2, **(1995)**, pp.263-286.
[3] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems", Proc. of the 13th Annual Conf. on Computational Learning Theory, **(2000)**, pp. 35-46.
[4] W. Utschick and W. Weichselberger, "Stochastic organization of output codes in multiclass learning problems", Neural Computing, vol. 13, no. 5, **(2001)**, pp. 1065-1102.
[5] K. Ludmila I, "Using diversity measures for generating error-correcting output codes in classifier ensembles", Pattern Recognition Letters, vol. 26, no. 1, **(2005)**, pp. 83-90.
[6] Y. Jiang, Q. Zhao and X. Yang, "A Search Coding Method and Its Application in Supervised Classification", Journal of Software, (In Chinese), vol. 16, no. 06, **(2005)**, pp. 1081-1089.
[7] F. Masulli and G. Valentini, "An experimental analysis of the dependence among codeword bit errors in ECOC learning machines", Neuro computing, vol. 57, **(2004)**, pp. 189-214.
[8] X. Jiantao and H. Mingyi, "Multiclass Classification Using Support Vector Machines (SVMs) Combined with Error-Correcting Codes (ECCs)", Journal of Northwestern Polytechnical University, (In Chinese), vol. 21, no. 4, **(2003)**, pp. 443-448.

Authors

Zhigang Yan received B.Sc. degree from China University of Mining and Technology in 1997 and Ph.D. degree from China University of Mining and Technology in 2007. He is currently a associate professor at faculty of China University of Mining and Technology, China. His field of interest is spatio-temporal data mining and knowledge discovering.

Yuanxuan Yang, Master student, received B.Sc. degree in Geographic Information System in 2013 from China University of Mining and Technology. Now he study in China University of Mining and Technology, supervised by Zhigang Yan.