Computers and Mathematics with Applications 57 (2009) 1908–1914

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A novel pattern recognition algorithm: Combining ART network with SVM to reconstruct a multi-class classifier

Anna Wang *, Wenjing Yuan, Junfang Liu, Zhiguo Yu, Hua Li

College of Information Science and Engineering, Northeastern University, 110004, Shenyang, China

ARTICLE INFO

Keywords:
ART network
Fault diagnosis
One-against-one
Multiclassification
SVM

ABSTRACT

Based on the principle of one-against-one support vector machines (SVMs) multi-class classification algorithm, this paper proposes an extended SVMs method which couples adaptive resonance theory (ART) network to reconstruct a multi-class classifier. Different coupling strategies to reconstruct a multi-class classifier from binary SVM classifiers are compared with application to fault diagnosis of transmission line. Majority voting, a mixture matrix and self-organizing map (SOM) network are compared in reconstructing the global classification decision. In order to evaluate the method's efficiency, one-against-all, decision directed acyclic graph (DDAG) and decision-tree (DT) algorithm based SVM are compared too. The comparison is done with simulations and the best method is validated with experimental data.

1. Introduction

Support vector machines (SVMs) classification is a modern machine learning method that has given superior results in various classification and pattern recognition problems. But SVM is the classifier that is originally designed for binary classification, so different coupling strategies to reconstruct a multi-class classifier from binary SVM classifiers are generally studied. Some multi-class classifiers [1] are commonly used, such as one-against-one (OAO), one-against-all (OAA), decision directed acyclic graph (DDAG) and decision-tree (DT) based SVM. However OAA algorithm has some unclassifiable regions; OAA-FSVM (fuzzy support vector machine) [2] is provided. Based on the analysis of the structure and the classification performance of the DTSVM [3], a separable measure has been defined. Based on the analysis of the structure and the shortage of OAO-SVM multi-class classifier, this paper proposes an extended OAO-SVM algorithm. Experimental results show the effectiveness of our scheme; experimental results also present the classification performance of this method comparing with other multi-class classification methods. The remainder of the paper is organized as follows. The basic theory on SVM is shown in Section 2. The multi-class classifier of different algorithms is introduced in Section 3. The extended OAO-SVM algorithm is introduced in Section 4. Section 5 presents the experimental analysis. Then the conclusion is given in Section 6.

2. SVM theory

SVM is a relatively new computational learning method based on the statistical learning theory presented by Vapnik [4]. In SVM, original input space is mapped into a high-dimensional dot product space called a feature space, and in the feature space the optimal hyperplane is determined to maximize the generalization ability of the classifier. The optimal hyperplane is found by exploiting the optimization theory, and respecting insights provided by the statistical learning theory.

Let n-dimensional input x_i ($i = 1, 2, \ldots, l$, l is the number of samples) belong to Class 1 or Class 2 and associated labels $y_i = 1$ be for Class 1 and $y_i = -1$ for Class 2, respectively. For linearly separable data, we can determine a hyperplane

* Corresponding author.
E-mail address: wanganna@mail.neu.edu.cn (A. Wang).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.10.052

$f(x) = 0$ that separates the data.

$$f(x) = \omega \cdot x + b = \sum_{i=1}^{n} \omega_i x_i + b = 0 \tag{2.1}$$

where ω is an n-dimensional vector and b is a scalar. The vector ω and the scalar b determine the position of the separating hyperplane. Function sgn $(f(x))$ is also called the decision function. A distinctly separating hyperplane satisfies the constraints:

$$y_i(x_i \cdot \omega + b) - 1 \geq 0 \Leftrightarrow \begin{cases} f(x_i) = x_i \cdot \omega + b \geq 1 & y_i = +1 \\ f(x_i) = x_i \cdot \omega + b \leq -1 & y_i = -1. \end{cases} \tag{2.2}$$

The separating hyperplane that creates the maximum margin is called the optimal separating hyperplane. Taking into account the noise with slack variables ξ_i and error penalty C, the optimal hyperplane can be found by solving the following convex quadratic optimization problem. Minimize:

$$\phi(\omega, \xi) = 1/2(\omega \cdot \omega) + C\left(\sum_{i=1}^{l} \xi_i\right) \tag{2.3}$$

subject to:

$$y_i[(x_i \cdot \omega) + b] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l \tag{2.4}$$

where ξ_i is measuring the distance between the margin and the example x_i lying on the wrong side of the margin. The calculations can be simplified by converting the problem with Kuhn–Tucker conditions into equivalent Lagrange dual problem.

$$V(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \tag{2.5}$$

subject to:

$$\sum_{i=1}^{l} y_i \alpha_i = 0, \quad C \geq \alpha \geq 0, i = 1, 2, \ldots, l. \tag{2.6}$$

The function $K(x_i \cdot x_j)$ that returns a dot product of the feature space mappings of original data points is called a kernel function. Three common kernel functions:

Polynomial kernel function: $K(x, x_i) = [(x \cdot x_i) + 1]^q$; Radial bass kernel function (RBF): $K(x, x_i) = \mathrm{e}^{-\frac{\|x - x_i\|}{2\sigma^2}}$; Sigmoid kernel function: $K(x, x_i) = \tan h(v(x \cdot x_i) + C)$.

The number of variables of the dual problem is the number of training data. Let us denote the optimal solution of the dual problem with $\alpha^*$and ω^*. According to the Karush–Kuhn–Tucker theorem, the equality condition holds for the training input–output pair (x_i, y_i) only if the associated α^* is not 0. In this case the training example x_i is a support vector (SV). The number of SVs is considerably lower than the number of training samples making SVM computationally very efficient.

The value of the optimal bias $b^*$is found from:

$$b^* = -\frac{1}{2} \sum_{SVs} y_i \alpha_i^* (K(x_i, x_j) + K(x_i, x_k)) \tag{2.7}$$

where x_j and x_k are arbitrary SVs for class 1 and class 2, respectively. The final decision function will be given by sgn $(f(x))$:

$$f(x) = \sum_{i,j=1} y_i \alpha_i^* K(x_i \cdot x_j) + b^*. \tag{2.8}$$

3. The multi-class classification of OAO-SVM

A solution is to decompose a multi-class problem to several two-class problems, train classifiers to solve these problems, and then reconstruct the solution of the multi-class problem from outputs of the classifiers, such as OAO-SVM. In the binary decomposition, $K(K-1)/2$ binary classifiers are built, each separating one class from another ignoring all the other classes. Several coupling strategies [5] are studied here to combine binary classifiers to find the global solution to the K-class problem.

3.1. Majority voting

An intuitive approach to get the global classification result is to consider outputs of the binary classifiers as votes and select the class that gets most votes. Assuming x is an input vector to be classified, and the classes are mutually exclusive, and the posteriori probabilities are $p_i = p(x \in class\, i)$, and the classifier $f_{ij}(x)$ discriminating between class i and j computes an estimate p_{ij} of the probability: $p_{ij} = p(x \in class\, i \bigcup class\, j) = \frac{p_i}{p_i + p_j}$.

The final classification decision is determined by: $F = \arg\max_{1\le i\le k} \sum_{j\neq i} \langle p_{ij} > 0.5\rangle$, here $\langle\cdot\rangle$ is defined by:

$$\langle z\rangle = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

The outputs of the classifiers cannot be interpreted as pure probabilities, like presented above. But the bigger the output classifier the more likely the sample will belong to the positive class and vice versa. Majority voting schemes are appropriate coupling strategies to build a multi-class classifier from SVMs.

3.2. Mixture matrix coupling

Applying majority voting methods, the vote counting takes into account the outputs of all binary classifiers with similar weights, without considering their significance. However, possible redundancy of some binary classifiers may be considered with the so-called mixture matrix. Let $f(x) = [f_1(x), f_2(x), \ldots, f_n(x)]$, where n is the number of the binary classifiers, and $f_i(x)$ is the output of i-classifier. It is assumed that a sized mixture matrix A [5] can be found, so that $\widetilde{f}(x) = f(x)A + e$, where k is the number of the classes, and the vector e represents noise. The mixture matrix is estimated to emphasize outputs of the classifiers that have strong correlation with a correct classification decision.

With this approach, outputs of the classifiers are linearly coupled with the mixture matrix created in the training phase to minimize the error between the correct class decision and the linear combination of outputs of the binary classifiers. Compared to the majority voting approach, this approach weights the votes given by classifiers so that the classification is more likely correct.

4. The extended OAO-SVM method

When the outputs of binary classifiers approach zero and the original OAO-SVM method applying majority voting strategy gets same votes, the classification results will be fallibility. This paper couples ART network and also couples SOM network [6] to reconstruct the multi-class classifier from binary classifiers. The performances are evaluated by the experiments on transmission line fault diagnose.

4.1. ART network

Adaptive Resonance Theory (ART) was invented by Stephen Grossberg in 1976. There are various unsupervised ART algorithms such as ART-1, ART-2, ART-3, and Fuzzy ART. In unsupervised ART nets, the input patterns may be presented several times and in any order. Each time a pattern is presented, an appropriate cluster unit is chosen, and related cluster weights are adjusted to let the cluster unit learn the pattern. Choosing a cluster is based on the relative similarity of an input pattern to the weight vector for a cluster unit rather than the absolute difference between the vectors. As in the most cases of clustering nets, the weights on a cluster unit may be considered as an exemplar for the patterns placed on that cluster. ART nets are designed to allow the user to control the degree of similarity of patterns placed on the same cluster through tuning the vigilance parameter. The vigilance parameter can be used to determine the proper number of clusters in ART nets, in order to reduce the probability of merging different types of clusters to the same cluster. Moreover, ART nets [7] have two other main characteristics: stability that means a pattern does not oscillate among different cluster units at different stages of training, and plasticity that means ART nets are able to learn a new pattern equally well at any stage of learning. Stability and plasticity of ART nets and the capability of clustering input patterns based on the user controlled similarity between them, made these nets more appropriate for transmission line diagnose rather than most of the other types of unsupervised neural nets. ART-1 is aimed to cluster binary inputs vectors. In the following sections, we discuss the practical comparison of the classifiers based on ART-1 with the SOM classifier as well as their capability in this application from various aspects.

In original algorithm of ART1 network, we calculate the inputs of every neuron as follows:

$$s_j = \sum_{i=1}^{n} w_{ij} \cdot x_i^k. \tag{4.1}$$

We define this formula as 0-matching. We can see that there is no effect on neuron's output unless the input feature is nonzero. When we calculate the degree of matching, we also only calculate the nonzero features. If we can calculate 0 feature weights, we can also calculate 1 feature weights in calculating matching degree. Then we calculate $s_j = \sum_{i=1}^{n} w_{ij} \cdot (1 - x_i^k)$. We define this formula as 1-matching. From the following experiments, because of processing the reversed information, this modification of the algorithm has lower ratio of misinformation. (If there is one fault, the classifier cannot detect it, and then misinformation is defined.)

4.2. The combination with artificial neural network (ANN)

Referring to Fig. 1, the multi-class classifier with the coupling of SVM and ART net is discussed in the following topics under the classifier's structure and constructional procedure respectively. The study focuses on the use of ART-1 net for coupling the outputs of binary classifiers. Assuming n-class classification problem, the number of the binary classifiers is

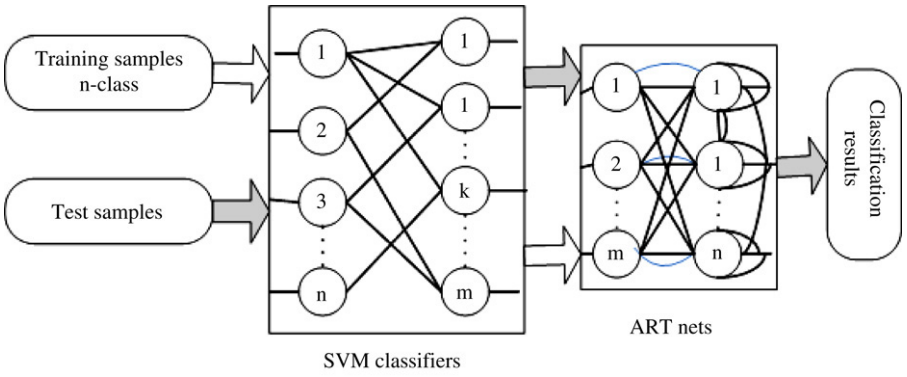

Fig. 1. The structure of multi-class classifier combined SVM with ART-1 net.

$m = n(n-1)/2$, and the number of the input-layer neurons is m, and the number of the output-layer neurons is n. From the following experiments, this method gives results very close to SVM of majority voting. Because the binary classifiers' outputs ($\mathrm{sgn}(f(x))$) are 0 or 1, we put them to ART-1 net without considering their significance. Whereas the comments above, we apply SOM net to couple the binary classifiers' outputs ($f(x)$ not $\mathrm{sgn}(f(x))$) directly. Although this method has a little higher efficiency, the time of training and testing phases is more than the SVM algorithm coupling ART-1 net.

4.3. Review of the extended OAO-SVM method

In order to couple the merits of the two improved methods above, this paper proposes an extended OAO-SVM method. Fig. 2 shows the classifying phase of the extended OAO-SVM, which includes three stages. One is feature extraction, other is training phase, and another is testing phase. Considering the significance of the binary classifiers' outputs ($f(x)$) [8], we can import a weight vector v which is trained by back propagation (BP) network. BP net is only validated in the training phase. Once the vector v is trained, the BP net is isolated in the testing process. The time of training process is more than the two improved methods above, but testing time is close to the first one. Training process can complete offline, so this method can also be applied to diagnose online. Fig. 3 expresses the structure of the extended OAO-SVM multi-class classification.

4.4. The algorithm of the extended OAO-SVM classifier

Step 1. We get a set of training samples, $x_k = (x_1^k, x_2^k, \ldots, x_t^k)$, $k = 1, 2, \ldots, p$, p and t are the number of input samples and its dimension, respectively. The expectable output vector is $y_k = (y_1, y_2, \ldots, y_m)$, where $y_i \in 0, 1$, m is the number of binary classifiers;

Step 2. Under the restriction conditions of formula (2.6) above, we can get α_k^* through solving the maximum of the formula (2.5) above;

Step 3. Calculate expressions: $V^* = \sum_{k=1}^{p} \alpha_k^* y_k x_k$ and b_k^* by formula (2.7);

Step 4. Select an appropriate kernel function (RBF) and calculate the formula (2.8) and get the results of qth binary classifiers: $f_q(x_k)$, $q = 1, 2, \ldots, m$. The total number of the binary classifiers is m;

Step 5. Switch to Step 1. until all samples are chosen.

Step 6. Let $O_i^k = (0, 0, \ldots, 1, 0, 0)$, O_i^k is the expectable output of the k-class. n is the number of the classes. Let all samples' outputs of binary classifiers $f(x)$ to BP net, weight vector v (m-dimension) is obtained by training. The BP net includes m-neuron of the input-layer and n-neuron of the output-layer.

Step 7. Considering the different significance of every classifier, we calculate vector $z^k = \mathrm{sgn}(f(x_k) \cdot v)$;

Step 8. Initialize inner star weights and outer star weights: $w_{qj}(0) = 1/(n+1)$, $t_{qj}(0) = 1$, where $j = 1, 2, \ldots, m$;

Step 9. Let the outputs of classifiers z^k to ART1 network;

Step 10. Calculate the inputs of the output-layer neuron of ART1 network: $s_j = w_j \cdot z^k = \sum_{q=1}^{m} w_{qj} \cdot (1 - z_q^k)$, if we apply 0-matching, we calculate $s_j = \sum_{q=1}^{m} w_{qj} \cdot z_q^k$;

Step 11. Select optimal classification result: $s_{j^*} = \max\{s_j\}$, let the output of neuron j^* is 1, the others are 0;

Step 12. Calculate the matching degree of and outer star weights, and make some decisions. $|z_q'| = \sum_{q=1}^{m}(1 - z_q^k)$ and $|T_{j^*} \cdot z_q'| = \sum_{q=1}^{m} t_{j^*q}(1 - z_q^k)$. If $|T_{j^*} \cdot z_q'|/|z_q'| > \rho$, it matches successfully and switches to Step 14, or else switches to Step 13;

Step 13. Cancel classification results and restore the neuron j^*. Neuron j^* will be excluded in the following classification process. Then switch to Step 10. If we have chosen all used neurons, there is no one satisfying the formula: $|T_{j^*} \cdot z_q'|/|z_q'| > \rho$. We must add a new neuron as the classification result and switch to Step 14;

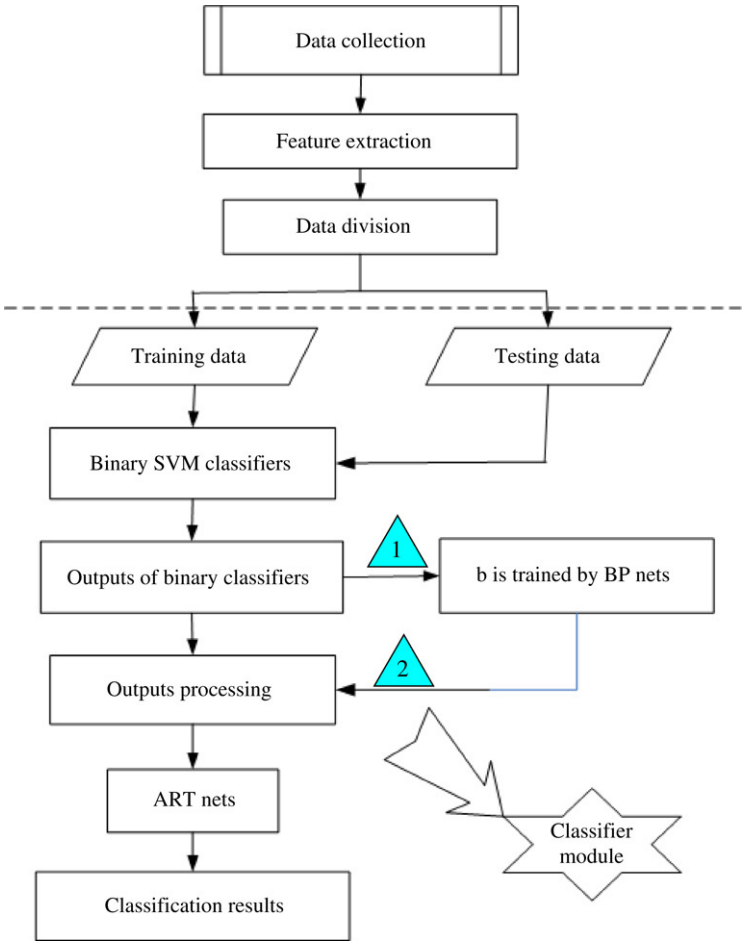

Fig. 2. Review of the extended OAO-SVM classifier.

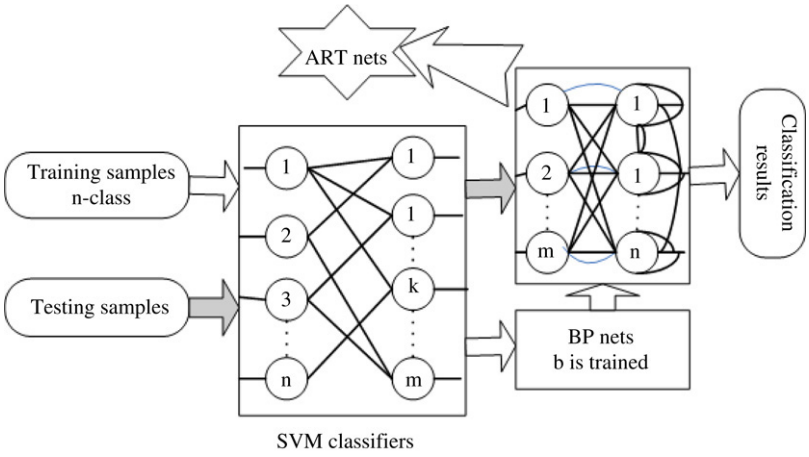

Fig. 3. The structure of the extended OAO-SVM classifier.

Step 14. Rectify connection weights: $w_{qj^*}(t+1) = \frac{t_{j^*q}(t)\cdot z_q^k}{0.5+\sum_{q=1}^{m} t_{j^*q}\cdot z_q^k}$

$$t_{j^*q}(t+1) = t_{j^*q}(t)\cdot z_q^k. \tag{4.2}$$

Step 15. Add all neurons that restored in Step 13 to the array of classification, then return Step 9 to classify the outputs of binary classifiers from next samples. The learning stage is over until all of the training samples are completed.

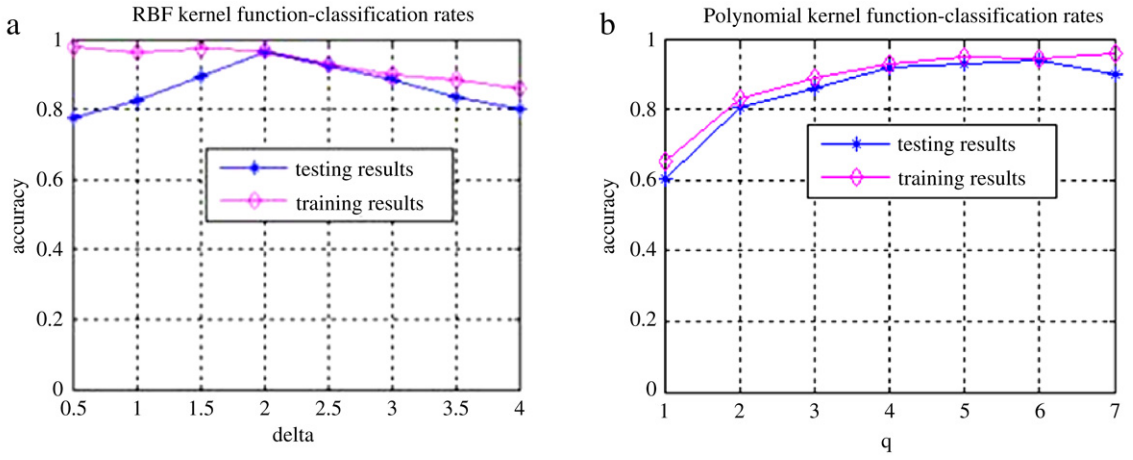

Fig. 4. (a) The classification rates of RBF kernel function. (b) The classification rates of polynomial kernel function.

Table 1
The accuracies of four multi-class classification methods of SVM.

Methods	The number of SVs	Training accuracy (%)	Testing accuracy (%)
OAA-SVM	48	90.50	88.33
DDGA-SVM	37	94.50	93.33
DT-SVM	31	96.50	95.00
EOAO-SVM	25	97.50	97.00

Table 2
The accuracies of different strategies for coupling binary classifiers of SVM.

Strategies	Training time (s)	Training accuracy (%)	Testing time (s)	Testing accuracy (%)
Majority voting	9.64	96.50	0.2810	94.00
SOM	13.21	94.50	0.3270	94.33
ART-1	12.58	96.00	0.2930	94.67
ART-1 with weigh	16.89	98.00	0.2990	97.67

5. Experimental results

Different types of faults are created on the power system network [9,10] such as line–ground (LG), line–line–ground (LLG), line–line (LL), and line–line–line–ground (LLLG). The data collected for voltage and current signals are real-time at the relaying end. It is found that the maximum change in standard deviation information for identifying the fault current and voltage pattern for a fundamental frequency of 50 Hz. In the proposed scheme, the features extracted form fault current and voltage signals are used to train and test the extended OAO-SVM (EOAO-SVM) for fault classification. The faulted voltage and current signal are processed through wavelet transform.

The SVM is trained with 200 sets and tested for 300 data sets. Kernel function of this method is selected by experiments. Fig. 4 presents the classification rates of two kernel functions. For polynomial kernel function, the highest points of training and testing accuracy are $q = 5$. For RBF kernel function, the highest points are $\sigma = 2.0$. The method of this paper proposed has higher efficiency for selecting RBF.

Table 1 shows the accuracies and the number of support vectors generated by different multi-class classification methods, respectively. It can be seen that the normal multi-class classifiers have generated more support vectors than the extended OAO-SVM classifiers. Since Vapnik showed that the number of support vectors is proportional to the generalization error of the classifier [4], so that the EOAO-SVM classifiers have a better capability of generalization.

Table 2 shows the efficiency of the different strategies for constructing multi-class classifiers. It can be seen that the strategy of majority voting applying for multi-class classification is superior to training time. However, its testing accuracy can't be accepted. The method, applying for ART1 without considering the significant of every binary classifier, is lower in testing accuracy. Thus we improve this method and give a weight to every binary classifier. From the experiments, it can be seen that this improved method have more accuracy. Though the method which applies strategies of ART1 1-matching with weight costs more time in training process, its testing time is close to any other methods. So we can detect fault online.

The fault classification rates for different kinds of faults are given in Table 3. In the table Average classification rates are calculated from the total number of the ith class samples, $\frac{1}{3}(\frac{m_1}{n_1} + \frac{m_2}{n_2} + \frac{m_3}{n_3})$, which is the number of samples classified correctly. The classification rate is 99.16% in case of LG fault which is the highest. The classification rates given by extended

Table 3
The classification rates for different transmission line faults.

Faults	Classes	Classification rates (%)	Average classification rates (%)
	AG	98.52	
LG	BG	100	99.16
	CG	97.89	
	ABG	96.35	
LLG	BCG	96.23	96.83
	CAG	97.48	
	AB	95.67	
LL	BC	97.52	96.82
	CA	97.28	
LLLG	ABCG	97.83	97.83

OAO-SVM shows that the misclassification is very less, which indicates robustness of extended OAO-SVM to classify faulty phases involved in the fault process. The coupling strategy of ART-1 with weight has two means. One is 0-matching which has lower misinformation. Experiments show that the rate of misinformation is 1.25%. The other is 1-matching whose misinformation is 3.20%.

6. Conclusions

This paper presents an extended multi-class classification method and applies it to detect the fault of transmission line. The current and voltage signals acquired from the relays were wavelet-transformed and features were extracted. The classification efficiency is much higher due to the coupling of SVM with ART and considering the significance of every binary classifier. The classification accuracy using the different constructing strategies is also presented in this paper. Thus, the trained network of extended OAO-SVM is capable of providing fast and precise fault classification.

References

[1] A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, Mechanical Systems and Signal Processing 21 (6) (2007) 2560–2574.
[2] T.-Y. Wang, H.-M. Chiang, Fuzzy support vector machine for multi-class text categorization, Information Processing and Management 43 (2007) 914–929.
[3] V. Sugumaran, V. Muralidharan, K.I. Ramachandran, Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing, Mechanical Systems and Signal Processing 21 (2007) 930–942.
[4] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1999.
[5] S. Pöyhönen, A. Arkkio, P. Jover, H. Hyötyniemi, Coupling pairwise support vector machines for fault classification, Control Engineering Practice 13 (2005) 759–769.
[6] B.-S. Yang, W.-W. Hwang, D.-J. Kim, A.C. Tan, Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines, Mechanical Systems and Signal Processing 19 (2005) 371–390.
[7] J.D. Martín-Guerrero, P.J.G. Lisboa, E. Soria-Olivas, A. Palomares, E. Balaguer, An approach based on the adaptive resonance theory for analysing the viability of recommender systems in a citizen web portal, Expert Systems with Applications 33 (2007) 743–753.
[8] D. Zhuang, B. Zhang, Q. Yang, J. Yan, Z. Chen, Y. Chen, Efficient text classification by weighted proximal SVM, in: Proceedings of the Fifth IEEE International Conference on Data Mining, 2005, pp. 538–545.
[9] N. Zhang, M. Kezunovic, A real time fault analysis tool for monitoring operation of transmission line protective relay, Electric Power Systems Research 77 (2007) 361–370.
[10] S.R. Samantaray, P.K. Dash, G. Panda, Distance relaying for transmission line using support vector machine and radial basis function neural network, Electrical Power and Energy Systems 29 (2007) 551–556.