Fusion of Multi-class Support Vector Machines for Fault Diagnosis

Zhonghui Hu, Yunze Cai, Xing He, and Xiaoming Xu

Abstract—**Data fusion strategies based on multi-class support vector machines are proposed. In the centralized scheme, the information from several sources is combined to construct an input space. In the distributed schemes, the input space is constructed corresponding to each information source and the multi-class support vector machine is used for modeling each source. The distributed data fusion strategies are applied to combine these multi-class support vector machine models. It is taken into account that a SVM classifier realizes classification by finding the optimal classification hyperplane with maximal margin. The proposed methods are demonstrated with the fault diagnosis of a diesel engine. The experimental results show that most of the proposed approaches can largely improve the diagnostic accuracy. The robustness of diagnosis is also improved.**

I. INTRODUCTION

FAULT diagnosis is still an ongoing research problem, for the machinery become more and more complex and the failure mode become more and more complicated, and the enormous improvements in the performance and cost of digital signal processing and communication devices in recent years [1]. The current approaches of fault diagnosis mainly rely on the classical techniques in the time or frequency domain, and statistical analysis [2]. The artificial intelligence method can often obtain a better performance.

It is difficult to diagnose more than one category of faults and some results obtained from fault specific method are not easy to interpret [3]. Shen et al. proposed a rough set theory based method that can diagnose more than one category of faults. One drawback of this method is that the rough set theory cannot be used to deal with the continuous attributes. The discretization method has to be used when continuous attributes exist. Because a prior knowledge about the attribute is difficult to obtain, it is hard to select an appropriate discretization method.

As a new generation learning system based on recent advances in statistical learning theory, Support vector machines (SVMs) have been established as one of the standard tools for machine learning and data mining [4]. It excludes the problem of local minima encountered in training neural networks. It is a promising theory for the application to fault diagnosis. SVMs are originally developed to solve binary classification problems. However, the discrimination between more than two categories is often required. Currently there are two types of approaches for multi-class SVM (MSVM) [5]. One is by constructing and combining several binary SVM classifiers while the other is by directly considering all data in one optimization formulation. Hsu et al. [5] indicated that the first approach is more suitable for practical use. In our methods of fault diagnosis, three typical methods of MSVM, "one-against-all," "one-against-one," and directed acyclic graph SVM (DAGSVM), are applied and evaluated respectively.

The information from several sources can be used in order to achieve higher classification accuracy and robustness [6, 7]. The conventional statistical pattern recognition methods, in most cases, are not appropriate in classification of multi-source data. For multi-source multi-class classification problem, the application of MSVMs is ongoing. In this paper the schemes using the data fusion techniques to combine MSVMs are proposed to solve the problem of multi-source multi-class classification.

One of the practical limitations in feature fusion is that the tremendous size of the feature space and the resulting heavy computational burden. To alleviate this problem, many researchers do not consider dependence among features from different sources, so that decisions are made individually based on signals from each source, and then combined together. This method is known as decision level fusion [8]. The feature level and the decision level are more practical in fusion, and the decision level fusion is most flexible. Therefore, the feature level and decision level fusion are mainly considered in this paper.

This paper is organized as follows. In Section II, the standard SVMs for binary classification is reviewed. We

Manuscript received September 12, 2004. This work was supported by by the "973" Program (2002cb312200), and partially supported by the "863" Program (2002AA412010), and the Natural Science Foundation of China (60174038).

Zhonghui Hu is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (Tel/Fax: +86.21.62826946; e-mail: huhzh@sjtu.edu.cn).

Yunze Cai is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: yzcai@sjtu.edu.cn).

Xing He is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: xhe@sjtu.edu.cn).

Xiaoming Xu is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: xmxu@sjtu.edu.cn).

introduce one-against-all, one-against-one, and DAGSVM methods in Section III. In Section IV, the fusion strategies for distributed fault diagnosis are discussed. Numerical experiments are given in Section V. Finally, conclusions are given in Section VI.

II. Support Vector Machines for Binary Classification

A comprehensive description of SVMs can refer to [4]. This section concisely describes the SVM theory. For the training data set $\{(x_i, y_i)\}_{i=1}^{l} \in R^n \times \{+1, -1\}$, where x_i represents condition attribute and y_i represents class attribute. Only the cases of linearly inseparable and nonlinearly inseparable are discussed in the following.

For the linearly inseparable case, the optimal classification hyperplane is obtained by solving

$$\begin{aligned} &\min\ J(W,\xi) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t.\ y_i\left[W \cdot x_i + b\right] \geq 1 - \xi_i, \\ &\quad \xi_i \geq 0, \quad i = 1,2,\cdots,l \end{aligned} \tag{1}$$

where C is the constant of capacity control and ξ_i is the slack factor that permits margin failure of corresponding x_i.

The optimization problem (1) can be rewritten as, applying the Lagrange optimization method and duality principle,

$$\begin{aligned} &\max\ M(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \langle x_i \cdot x_j \rangle + \sum_{i=1}^{l}\alpha_i \\ &s.t.\quad \sum_{i=1}^{l}\alpha_i y_i = 0, \\ &\quad \alpha_i \in [0, C],\ i = 1,2,\cdots,l \end{aligned} \tag{2}$$

By solving the problem (2), we can get the optimal hyperplane with maximal margin

$$f(x) = \sum_{sv}\alpha_i y_i \langle x \cdot x_i \rangle + b = 0 \tag{3}$$

Therefore, the decision function based on SVM for linear classification in the input space is

$$d(x) = \operatorname{sgn}\left[f(x)\right] = \operatorname{sgn}\left[\sum_{sv} y_i\alpha_i \langle x_i \cdot x \rangle + b\right] \tag{4}$$

For the nonlinearly inseparable case, the problem is transformed as that of linear classification in the space H by projecting the original data into a high dimensional space H using a nonlinear map $\Phi : R^n \to H$. The problem (1) can be transformed to the more general kernel version [9]

$$\begin{aligned} &\min\ J(W,\xi) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t.\ y_i\left[W \cdot \Phi(x_i) + b\right] \geq 1 - \xi_i, \\ &\quad \xi_i \geq 0, \quad i = 1,2,\cdots,l \end{aligned} \tag{5}$$

By introducing the kernel $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, the problem (5) can be rewritten as

$$\begin{aligned} &\max\ M(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{l}\alpha_i \\ &s.t.\quad \sum_{i=1}^{l}\alpha_i y_i = 0, \\ &\quad \alpha_i \in [0, C],\ i = 1,2,\cdots,l \end{aligned} \tag{6}$$

By solving the above problem, we can get the optimal hyperplane with maximal margin

$$f(x) = \sum_{sv}\alpha_i y_i K(x, x_i) + b = 0 \tag{7}$$

The decision function of SVMs in the input space is

$$d(x) = \operatorname{sgn}\left[f(x)\right] = \operatorname{sgn}\left[\sum_{sv} y_i\alpha_i K(x_i, x) + b\right] \tag{8}$$

III. Multi-class Support Vector Machines

Three approaches of creating MSVM by training and combining several binary-class SVM classifier, one-against-all, one-against-one, and DAGSVM, are discussed in this section [5].

Given the training data set $\{(x_i, y_i)\}_{i=1}^{l}$, where $x_i \in R^n$ represents condition attribute and $y_i \in \{1,\cdots,k\}$ is the class attribute of x_i. The objective of multi-class classification is to correctly discriminating these classes from each other.

The earliest used implementation for MSVM classification is probably the one-against-all method [5]. The MSVM using one-against-all strategy can be constructed by applying the following procedure [10]:

(1) Construct k binary SVM classifiers where $f_i(x), i = 1,\cdots,k$ separates training data of class i from the other training data ($\operatorname{sgn}[f_i(x)] = 1$, if instance x belongs to class i, $\operatorname{sgn}[f_i(x)] = -1$ otherwise).

(2) Construct the k-class MSVM classifier by choosing the class corresponding to the maximal value of functions $f_i(x),\ i = 1,\cdots,k$. The decision function is

$$d(x) = \arg\ \max\left\{f_1(x),\cdots,f_k(x)\right\} \tag{9}$$

The one-against-one method constructs classifiers where each one is trained on data from two classes. There are different methods for doing the future testing after all classifiers are constructed. Hsu and Lin [5] used the voting strategy of "Max-Wins". However, if two classes have identical votes, the decision is not sound. A modified testing strategy is proposed as below.

In our modified testing strategy, the "Max-Wins" strategy is firstly applied to test the one-against-one MSVM. If more than one class have the identical votes, the following strategy is used based on the results of the "Max-Wins" strategy.

Suppose there exist p classes of which the max votes are equal to m. Each of the p classes has m functions $f_{ij}(x)$, $i=1,\cdots,p$, $j=1,\cdots,m$. The class attribute of x is determined by the sum of maximal distances to the optimal classification hyperplane. Thus, the auxiliary decision function is

$$d(x)=\arg \max_{\substack{i\in\{h_1,\cdots,h_p\}\\ \{h_1,\cdots,h_p\}\subseteq\{1,\cdots,k\}}} \left\{\sum_{j=1}^{m}\left|f_{ij}(x)\right|\right\} \tag{10}$$

The training phase of DAGSVM [11] is the same as the one-against-one method. However, in the testing phase, the decision is obtained by using a rooted binary directed acyclic graph which has internal nodes and leaves. Each node is a binary SVM of the *i*th and *j*th classes. Given a test instance, starting at the root node, the binary decision function is evaluated. Then it moves to either left or right depending on the output value. Through a path, we reach a leaf node which indicates the predicted class. An advantage of using a DAG is that the generalization can be analyzed [5]. In addition, its testing time is less than the one-against-one method.

IV. Data Fusion Strategies

The synergistic use of overlapping and complementary data sources provides information that is otherwise not available from individual sources. Furthermore, multiple data sources can provide more robust performance due to the inherent redundancy [12]. The traditional architecture for fusion is centralized. Due to the advances of computing and communication, the distributed architecture becomes feasible. In this architecture the data from individual data sources in lower level nodes are processed and then the results are sent to higher level nodes to be combined. The distributed fusion architecture has the following advantages: lighter processing load at each fusion node; no need to maintain a large centralized database; lower communication load; higher robustness.

In this paper, the centralized and distributed architectures based on MSVMs are proposed. The proposed schemes make full use of the characteristic of MSVMs. One is that the MSVMs discussed in this paper are created by constructing and combining several binary SVM classifiers. The other is that the training of a binary SVM classifier is to obtain the optimal classification hyperplane with maximal margin. The centralized scheme is labeled as the fusion scheme I. The distributed schemes include the fusion scheme II, III, IV and V. In each of the proposed schemes, the aforementioned three types of MSVM are applied separately.

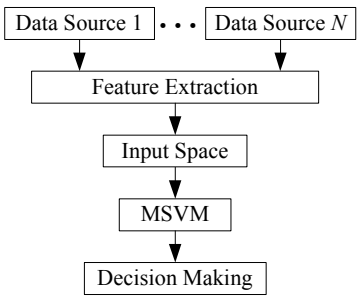

Fig. 1. Data fusion scheme I

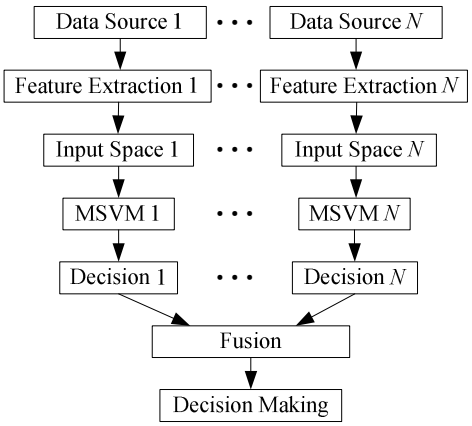

Fig. 2. Data fusion scheme II

The fusion scheme I is illustrated in Fig.1. The features of all the data sources are extracted and combined to form a single input space. Then the MSVM is trained and tested to create a decision maker.

In the fusion scheme II (Fig.2), the features of every data source are extracted and used to form an input space, respectively. Then the sub-MSVM decision makers are created. Suppose that the class attribute set is $CS=\{1,2,\cdots,k\}$. The final decision using the majority vote strategy is

$$\begin{aligned} &d(x)=\arg\max\{V_1,\ V_2,\cdots,V_k\},\\ &V_i=\sum_{j=1}^{N}\delta_{ij},\\ &\delta_{ij}=\begin{cases}1, & d_j=i\\ 0, & d_j\neq i\end{cases},\quad (\ \mathrm{i}=1,\cdots,k,\ j=1,\cdots,N) \end{aligned} \tag{11}$$

where $d(x)$ is the final decision function, V_i is the obtained votes of class i and d_j is the output of the *j*th MSVM trained by using the *j*th data source. It is possible that more than one class have the equal maximal votes. That is the limitation of this scheme. In this situation, the final decision can be decided by the following strategy.

For the one-against-all MSVM, suppose the max votes of p classes are equal to m. Thus, the *i*th class has m functions $f_{ij}(x)$, $i=1,\cdots,p$, $j=1,\cdots,m$. The class attribute of *x* is determined by the sum of maximal distances to the optimal classification hyperplane.

$$d(x)=\arg \max_{\substack{i\in\{h_1,\cdots,h_p\}\\ \{h_1,\cdots,h_p\}\subseteq\{1,\cdots,k\}}} \left\{\sum_{j=1}^{m} f_{ij}(x)\right\} \tag{12}$$

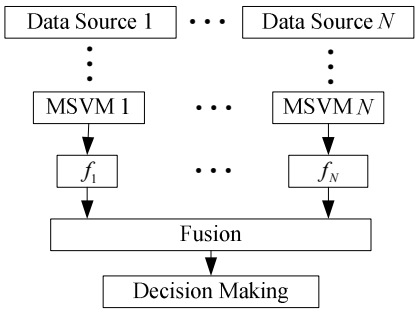

Fig. 3. Data fusion scheme III

For the one-against-one and DAGSVM, the decision function is

$$d(x) = \arg \max_{\substack{i \in \{h_1,\cdots,h_p\} \\ \{h_1,\cdots,h_p\} \subseteq \{1,\cdots,k\}}} \left\{ \sum_{j=1}^{m} \left| f_{ij}(x) \right| \right\} \tag{13}$$

In the fusion scheme III (Fig.3), the difference from the fusion scheme II is that the output of the *i*th MSVM is not a decision, but the maximal distance of the instance to the optimal classification hyperplane. For the one-against-all and DAGSVM, the function is defined as

$$f_i = \max\left\{ f_{i1}, f_{i2}, \cdots, f_{iN} \right\},\ i = 1, \cdots, k \tag{14}$$

For the DAGSVM, the function f_{iq} in (14) is given by

$$f_{ij} = f_d,$$
$$f_d = \begin{cases} 0, & i \neq r \\ |f_{ij}(x)|, & i = r \end{cases}, i = 1, \cdots, k,\ j = 1, \cdots, N \tag{15}$$

For the one-against-one MSVM, the function f_{ij} is

$$f_{ij} = \sum_{l=1}^{V_{ij}} \left| f_{ijl}(x) \right| \tag{16}$$

where the V_{ij} is the vote number of the *i*th class of the *j*th MSVM trained by using the *j*th data source. Thus, the final decision function is

$$d(x) = \max\left\{ f_1, f_2, \cdots, f_k \right\} \tag{17}$$

In the fusion scheme IV (Fig.4), the difference from the fusion scheme III is that the outputs of each binary SVM in the *i*th MSVM, the respective maximal distances of the instance to the optimal classification hyperplane, are used as features [13].

$$d(x) = \arg \max_{i \in \{1,\cdots,k\}} \left\{ \sum_{j=1}^{N} c_{ij} \left| f_{ij}(x) \right| \right\}, c_{ij} = \left| f_{ij} \right| \Big/ \sum_{\substack{i=1,\cdots,k \\ j=1,\cdots,N}} \left| f_{ij} \right| \tag{18}$$

For the one-against-one MSVM, there exists another fusion strategy illustrated by Fig.5. The decision function is given as follows.

$$d(x) = \arg\max\left\{ V_1,\ V_2, \cdots, V_k \right\}, V_i = \sum_{j=1}^{N} v_{ij}, \tag{19}$$

where v_{ij} is the vote number of the *i*th class of the *j*th MSVM trained by using the data from the *j*th data source.

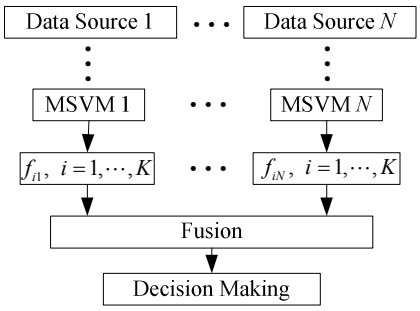

Fig. 4. Data fusion scheme IV

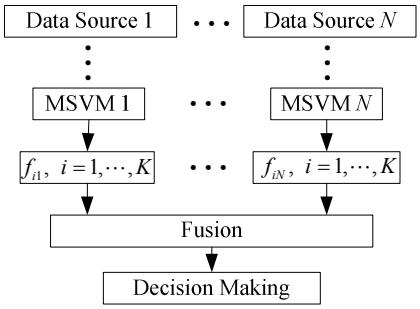

Fig. 5. Data fusion scheme V

V. Experimental Results

Rough set theory is used to diagnose the valve fault for a multi-cylinder diesel engine [14]. Four states are researched in [14]: Normal state; intake valve clearance is too small; intake valve clearance is too large; exhaust valve clearance is too large. Among these four states, three fault types are simulated in the intake valve and exhaust valve on the second cylinder head. Three sampling points are selected to collect vibration signals. They are the first cylinder head, the second cylinder head and another one at the centre of the piston stroke, on the surface of the cylinder block. Six features are extracted from the vibration signals. These features present the information contained in vibration signals both from the frequency domain and time domain. Thus, each instance in the dataset is composed of 18 condition attributes (six features from each sampling point) and one class attribute (four states). In the distributed schemes, the six features from one sampling point, adding the class attribute, form a dataset. Therefore, three datasets corresponding to three data sources are constructed.

The two-fold cross-validation test is used to show the effect of the rough set theory for fault diagnosis in [14]. The classification accuracy is listed in Table I. We can find that the average classification accuracy is 76.32%.

TABLE I
Classification Accuracy of Each Part
(training data: TRD; testing data: TD)

Data set	1st part: TRD 2nd part: TD	2nd part: TRD 1st part: TD
Accuracy (%)	78.95	73.68

TABLE II
COMPARISON OF THE EXPERIMENTAL RESULTS (%)
(D1, D2, D3: Data Source 1, 2, and 3; S.I, S.II, S.III, S.IV, S.V: Scheme I, II, III, IV, and V)

MSVM	Single Data Source			S.I	S.II	S.III	S.IV	S.V
	D1	D2	D3					
One-against-all	91.25	75.42	91.67	94.59	95.00	92.50	96.25	—
One-against-one	92.92	80.00	92.50	94.59	96.25	92.92	95.00	96.25
DAGSVM	92.92	80.00	92.50	94.59	95.00	87.92	94.17	—

The whole dataset is listed in [3]. It consists of 37 instances, among which 25 instances are used as training set and the rest are used as testing set. The choice of kernel and the regularizing parameter is determined via performance on a validation set. Eighty percent of the training set is used for training binary SVM classifiers and the rest 20% of the training set is used as validation set. The testing set is used to test the classification accuracy in different strategies. The experimental results are given in Table II. The data from every data source (sampling point) are used to train and test MSVM respectively. The classification accuracy of three types of MSVM are given in the D1, D2 and D3 columns. Scheme I is the centralized fusion strategy. All the Schemes from II to V are distributed strategies of data fusion. Scheme II is mainly based on max-distance strategy. Scheme III is mainly based on majority-vote strategy. Scheme IV is mainly based on max-weighted-distance strategy. Scheme V is also based on majority-vote strategy in some sense. Table II shows that all the classification accuracy of the approaches using fusion strategies outperforms that of other approaches without fusion strategies, except for Scheme III for DAGSVM. Scheme IV for one-against-all, as well as Scheme V for one-against-one, has the best accuracy.

VI. CONCLUSIONS

The centralized and distributed data fusion strategies based on MSVMs are proposed in this paper. They are evaluated by applying to fault diagnosis for diesel engine. Three methods constructing MSVMs by combining several binary SVM classifiers, one-against-one, one-against-all and DAGSVM, are mainly discussed. The proposed schemes make full use of the fact that the training of a binary SVM classifier is by finding the optimal classification hyperplane with maximal margin. When the data fusion strategy is not used, we can find that the one-against-one and DAGSVM methods have higher accuracy. In the data fusion schemes, the fusion schemes of II and III are most suitable for the one-to-one method, and the fusion scheme IV is most suitable for the one-against-all method. The improvement of accuracy is outstanding. The one-to-one method for fusion strategy IV also obtains high accuracy. In addition, the highest accuracy is also gained in the fusion strategy V for the one-to-one method. Our proposed methods can improve the robustness and accuracy of fault diagnosis. Furthermore, they can also be applied in other field.

REFERENCES

[1] M. Hajiaghajani, H. A. Toliyat, and I. M. S. Panahi, "Advanced Fault Diagnosis of a DC Motor," *IEEE Trans. Energy Conversion,* Vol. 19, no. 1, pp. 60-65, 2004.

[2] C. Chen and C. Mo, "A Method for Intelligent Fault Diagnosis of Rotating Machinery," *Digital Signal Processing*, Vol. 14, pp. 203-217, 2004.

[3] L.Shen, F. E. H. Tay, L. Qu and Y. Shen, "Fault diagnosis using Rough Sets Theory," *Computer in Industry*, Vol. 43, pp. 61-72, 2000.

[4] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and other kernel-based learning methods," Cambridge University Press. 2000.

[5] C. W. Hsu and C. J. Lin, "A Comparison of Methods for Multi-class Support Vector Machines," *IEEE Trans. Neural Networks*, Vol. 13, no. 2, pp. 415-425, 2002.

[6] D. L. Hall and J. Llinas, "An Introduction to Multisensor Data Fusion," *Proceedings of the IEEE*, Vol. 85, no. 1, pp. 6-23, 1997.

[7] J. A. Benediktsson, J. R. Sveinsson, O. K. Ersoy and P. H. Swain, "Parallel consensual neural networks," *IEEE Trans. Neural Networks*, Vol. 8, no. 1, pp. 54-64, 1997.

[8] H. Pan, Z. P. Liang, T. J. Anastasio and T. S. Huang, "A Hybrid NN-Bayesian Architecture for Information Fusion," *Image Processing, ICIP 98*, vol. 1, pp. 368–371, 1998.

[9] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol. 2, no. 2, pp. 121-167, 1998.

[10] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[11] J. C. Platt, N. Cristianini and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, Vol. 12, pp. 547–553, 2000.

[12] M. E. Liggins, C. Y. Chong, I. Kadar, M. G. Alford, et al., "Distributed Fusion Architectures and Algorithms for Target Tracking," *Proceedings of the IEEE*, Vol. 85, no. 1, pp. 95-107, 1997.

[13] W. Yan, H. Shao and X. Wang, "Parallel Decision Models Based on Support Vector Machines and their Application to Distributed Fault Diagnosis," *Proceedings of the American Control Conference*, Vol. 2, pp. 1770-1775, 2003.

[14] F. E. H. Tay and L. Shen, "Fault diagnosis based on Rough Set Theory," *Engineering Application of Artificial Intelligence*, Vol. 16, pp.39-43, 2003.