DEMPSTER-SHAFER THEORY BASED MULTI-CLASS SUPPORT VECTOR MACHINES AND THEIR APPLICATIONS

Zhonghui Hu, Rupo Yin, Yuangui Li, Xiaoming Xu

Department of Shanghai Jiaotong University, Shanghai, 20030, P. R. China
Tel/Fax:+86.21.62826946; Email: huhzh@sjtu.edu.cn

Abstract: How to extend standard support vector machines to solve multi-class classification problem and yield the outputs in the frame of Dempster-Shafer theory is useful. The multi-class probability support vector machine is proposed, firstly. The Dempster-Shafer theory based multi-class support vector machine is designed by constructing probability support vector machines for binary classification using one-against-all strategy and then combining them using Dempster-Shafer theory. Our proposed method is applied to fault diagnosis for a diesel engine. The experimental results show our proposed method obtains a comparable performance with that of standard multi-class support vector machines. Furthermore, the uncertainty can also be evaluated. *Copyright © 2005 IFAC*

Keywords: Machine learning; Classifiers; Classification; Fault diagnosis; Diesel engines

1. INTRODUCTION

Support vector machines (SVMs), deriving from statistical learning theory and VC-dimension theory, have been widely used in many fields and show good performance (Vapnik, 1995). It is originally developed to solve binary classification problems. However, in real world problems, the discrimination between more than two categories is required. How to extend the standard SVM to solve multi-class problem is an ongoing research problem. Currently there are two types of approaches for multi-class SVM (MSVM) (Hsu and Lin, 2002). One is by constructing and combining several binary classifiers while the other is by directly considering all data in one optimization formulation. The one-against-all strategy is the standard method for constructing MSVM (Vapnik, 1995; Platt, *et al.*, 2000).

Generally, a posterior probability is convenient for post-processing. However, both standard SVM and MSVM classifiers do not provide such probabilities. Platt (1999) describes a method for fitting a sigmoid that maps SVM outputs to posterior probabilities, while still maintaining their sparseness. However, the probability SVM (PSVM) can still not be directly used to solve multi-class classification problem. Therefore, in this paper, we extend the standard MSVM method, one-against-all, to multi-class probability SVM (MPSVM) method. Thus, all the outputs of MPSVM are presented as posterior probabilities. The final classification output of MPSVM is the class that corresponds to the PSVM with the highest probability output value.

The Dempster-Shafer theory is a general extension of Bayesian theory, which can robustly deal with incomplete data (Rakar and BalleÂ, 1999). Additionally, it allows assigning measures of probability to focal elements, and attaching probability to the frame of discernment. How to extend the support vector machine to yield the outputs in the frame of Dempster-Shafer theory is very useful. For example, the outputs of this type of SVM can directly be combined by using Dempster-Shafer theory, and more useful classification information can be obtained. Furthermore, a sound basis is constructed for the application of support vector machines in some other fields, such as information fusion. By designing the basic probability assignment (*bpa*) function according to all the outputs and performances of PSVMs consisting in the MPSVM and then combining all the evidences, the Dempster-Shafer theory based multi-class support vector machine (DSMSVM) is constructed. Different from the standard MSVM and MPSVM, the performance of individual PSVMs for binary classification in DSMSVM is also taken into account. Our proposed method is applied to fault diagnosis for a diesel engine. The experimental results show the proposed method obtains a comparable performance with the standard MSVM and MPSVM. Furthermore, the belief, plausibility, belief interval and ignorance about classes can also be provided.

This paper is organized as follows. In Section 2, the standard SVMs and PSVMs are reviewed. The standard MSVM is also introduced. Then, the MPSVM method is proposed. Section 3 describes the Dempster-Shafer evidence theory. In Section 4, the DSMSVM is proposed and a numerical example is given. In Section 5, by applying these methods to fault diagnosis for a diesel engine, comparison of different MSVMs is provided. Finally, conclusions are given in Section 6.

2. MULTI-CLASS PROBABILITY SUPPORT VECTOR MACHINES

It is generally simpler to construct classifier theory and algorithms for two classes than for more than two classes (Platt, *et al.*, 2000). Therefore, it is a better strategy to combine many two-class classifiers into a multi-class classifier.

In this section, we first introduce the standard SVM for binary classification. Second, the PSVM is introduced. Third, the standard MSVM method is reviewed. Finally, based on the PSVM and standard MSVM, the MPSVM method is proposed.

2.1 Support Vector Machines for binary classification

For the training data set $\left\{(x_i, y_i)\right\}_{i=1}^{l} \in R^n \times \{+1,-1\}$, where x_i represents condition attribute and y_i represents class attribute, SVMs optimize the classification boundary by separating the data with the maximal margin hyperplane. The practical data are usually inseparable. These cases are discussed as below.

For linearly inseparable case, the optimal classification hyperplane can be obtained by solving the optimization problem

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

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

According to the Lagrange optimization method and duality principle, the optimization problem (1) can be rewritten as follows.

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

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

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

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

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

For nonlinearly inseparable case, the original data are projected into a certain high dimensional Euclidean space H by a nonlinear map $\Phi: R^n \to H$, so that the problem of nonlinear classification is transferred into that of linear classification in the space H. By introducing the kernel function $K(x_i, x_j) = \left\langle \Phi(x_i), \Phi(x_j)\right\rangle$, it is not necessary to explicitly know $\Phi(\bullet)$ (Burges, 1998). Hence, the optimization problem (1) can be moved directly to the more general kernel version

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

The problem (5) can be rewritten as follows

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

By solving (6), we can get the optimal classification hyperplane

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

and the decision function that separates training vectors into two classes in the input space

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

2.2 Probability Support Vector Machines for Binary Classification

The PSVM method is provided by Platt where a sigmoid is trained to map standard SVM outputs to posterior probabilities, and the sparseness of SVM is still retained (Platt, 1999).

Let the real number output of a standard SVM, the distance from the unknown example x to the optimal classification hyperplane, be f (or $f(x)$). Instead of estimating the class-conditional density $p(f \mid y)$, a parametric model is used to fit the posterior $P(y=1 \mid f)$ directly. The parameters A and B of the model are adapted to give the best probability outputs.

$$P(y=1 \mid f)=\frac{1}{1+\exp(Af+B)} \tag{9}$$

The parameters A and B are found by minimizing the negative log likelihood of the training data, which is a cross-entropy error function

$$\min \ -\sum_{\mathrm{i}} \mathrm{t}_{\mathrm{i}} \log(p_i)+(1-t_i)\log(1-p_i) \tag{10}$$

where

$$p_i=\frac{1}{1+\exp(Af_i+B)} \tag{11}$$

$$t_i=\frac{y_i+1}{2} \tag{12}$$

Based on equation (12) and the training set (f_i, y_i), a new training set (f_i, t_i) is defined. The optimization problem (10) is solved by using the model-trust minimization algorithm for robustness (Gill, *et al.*, 1981).

A method of preventing overfitting in training sigmoid is used by Platt (1999). The out-of-sample data is modeled with the same empirical density as the sigmoid training data, but with a finite probability of opposite label. When a positive example is observed at a value f_i, we do not use $t_i=1$, but assume that there is a finite chance of opposite label at the same f_i in the out-of-sample data. Therefore, a value of $t_i=1-\varepsilon_+$ will be used, for some ε_+.

The probability of correct label can be derived using Bayesian rule. Suppose N_+ positive examples are observed. The maximum a posteriori (MAP) estimate for the target probability of positive examples is

$$t_+=\frac{N_++1}{N_++2} \tag{13}$$

Similarly, if there exist N_- negative examples, the *MAP* estimate for the target probability of negative examples is

$$t_-=\frac{1}{N_-+2} \tag{14}$$

Hence, the training set for sigmoid fit is

$$(f_i, t'_i),\ t_i^{'}=\begin{cases} t_+,\ t_i=1 \\ t_-,\ t_i=0 \end{cases},\ i=1,\cdots,l \tag{15}$$

Instead of $\{0,1\}$, these non-binary targets are used for all the data in the sigmoid fit. Moreover, the non-binary targets will converge to $\{0,1\}$ when the training set size approaches infinity, which recovers the maximum likelihood sigmoid fit.

Thus, by training sigmoid using the modified training set, the output of PSVM for unknown example x is

$$p(x)=\frac{1}{1+\exp(A\sum_{SV} a_i y_i K(x,x_i)+b+B)} \tag{16}$$

and the decision function of PSVM is

$$d(x)=\begin{cases} 1,\ p(x)\geq 0.5 \\ -1,\ p(x)<0.5 \end{cases} \tag{17}$$

2.3 Standard Multi-class Support Vector Machines

For a *K*-class classification problem, the standard MSVM method is to construct K standard SVMs for binary classification and then combining them (Platt, *et al.*, 2000; Vapnik, 1998). The *kth* SVM will be trained using all of the instances, among which the instances belonging to the *kth* class are labelled as positive, and all the other instances are labelled as negative. Therefore, the MSVM method trained in this way is also denoted as one-against-all or one-versus-rest method. The final classification output of standard MSVM is the class that corresponds to the SVM with the highest output value. Suppose the output of the *kth* SVM is $f_k(x)$, the final decision function is

$$d(x)=\arg\max\left\{f_1(x),\cdots\cdots,f_K(x)\right\} \tag{18}$$

2.4 Multi-class Probability Support Vector Machines

Given training data set $\left\{(x_i, y_i)\right\}_{i=1}^{l}$, 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 is to correctly discriminate these classes from each other. Based on the PSVM for binary classification and one-against-all strategy (Vapnik, 1998), the MPSVM can be constructed by applying the following procedure.

1) Construct K binary SVM classifiers where $f_k(x)$ ($k=1,\cdots,K$) separates training examples of the class k from the other training examples. The training set used for *kth* binary SVM is $\left\{(x_i, y_i')\right\}_{i=1}^{l}$ ($y_i'=1$, if $y_i=k$; $y_i'=-1$ otherwise).

2) Training the sigmoid using the modified training set $\{(f_i, t'_i)\}_1^l$, K binary PSVM classifiers with outputs $p_k(x)$, $k=1,\cdots,K$ are constructed.

3) Construct the K-class MPSVM classifier by choosing the class corresponding to the PSVM with the highest probability value among $p_k(x)$, $k=1,\cdots,K$. Therefore, the decision function is

$$d(x)=\arg\max\left\{p_1(x),\cdots\cdots,p_K(x)\right\} \tag{19}$$

3. DEMPSTER-SHAFER EVIDENCE THEORY

Dempster-Shafer evidence theory is regarded as a generalization of classic Bayesian theory. The attractive feature in evidence theory is that it not only directly takes into account what remains unknown, but also represents what is known precisely (Beynona, *et al.*, 2001; Guan and Bell, 1992).

Let $\Theta = \{h_1, h_2, \cdots, h_n\}$ be a frame of discernment, a function $m: 2^{\Theta} \to [0,\ 1]$ is called a basic probability assignment (*bpa*) if it satisfies

$$\begin{cases} m(\varnothing) = 0, \ (\varnothing \text{ - empty set}) \\ \sum_{X \in 2^{\Theta}} m(X) = 1 \end{cases} \tag{20}$$

where 2^{Θ} is the power set of Θ. Any subset X of the frame of discernment Θ with non-zero mass value is called a focal element and the mass function $m(X)$ represents the exact belief in the proposition corresponding to the subset X.

A belief function $bel:\ 2^{\Theta} \to [0,\ 1]$, derived from the mass function, is defined by

$$bel(A) = \sum_{X \subseteq A} m(X), \text{ for all } \mathrm{A} \subseteq \Theta \tag{21}$$

It represents the measure of the total belief lying in A and all subsets of A.

A plausibility function $pls: 2^{\Theta} \to [0,1]$ is defined as

$$pls(A) = 1 - bel(\overline{A}) = \sum_{A \cap X \neq \varnothing} m(X) \tag{22}$$

for all $A \subseteq \Theta$. Obviously $pls(A)$ denotes the extent to which we fail to disbelieve A. Furthermore, $bel(A) \leq pls(A)$ is always satisfied.

For a given subset *A*, a belief interval conveniently represents the information contained in the evidential functions $bel(A)$ and $pls(A)$, that is,

$$[bel(A),\ pls(A)] \tag{23}$$

Here the degree to which subset A remains plausible is measured by $pls(A)$. The difference between $pls(A)$ and $bel(A)$ represents the residual ignorance.

$$ignorance(A) = pls(A) - bel(A) \tag{24}$$

A method to combine the measures of evidence from N sources is also provided in Dempster-Shafer theory. The combined mass function $m_1 \oplus m_2 \oplus \cdots \oplus m_N:\ 2^{\Theta} \to [0,\ 1]$ is defined by

$$(m_1 \oplus m_2 \oplus \cdots \oplus m_N)(A) = \frac{1}{K_N} \sum_{X_1 \cap X_2 \cap \cdots \cap X_N = A} m_1(X_1) m_2(X_2) \cdots m_N(X_N) \tag{25}$$

where $x_1, x_2, \cdots, x_N$ are focal elements, and K_N is defined as

$$K_N = \sum_{X_1 \cap X_2 \cap \cdots \cap X_N \neq \varnothing} m_1(X_1) m_2(X_2) \cdots m_N(X_N) \tag{26}$$

where the constant $1/K_N$ measures the extent of conflict among these mass functions.

4. DEMPSTER-SHAFER THEORY BASED MULTI-CLASS SUPPORT VECTOR MACHINES

Suppose there exists a pattern space P containing K mutually exclusive classes, $\Gamma = \{1, 2, \cdots, K\}$ represents a class attribute set, which is also the frame of discernment Θ. Given training data set $\{(x_i, y_i)\}_{i=1}^{l}$, $x_i \in R^n$ represents condition attribute and $y_i = k \in \Gamma$ is the class attribute of x_i. Based on the MPSVM method, the DSMSVM can be constructed by using the following procedure.

1) Construct *K*-class MPSVM classifier consisting of K PSVM classifiers. The *i*th PSVM classifier is designed to discriminate examples of class *i* from all the other examples. The probability outputs and decision outputs of these PSVM classifiers are p_i and d_i, $i = 1, \cdots, K$, $p_i \in [0,1]$, $d_i \in \Theta$, respectively.

2) Obtain the performance of every PSVM classifier. The true classification accuracy of a classifier is not available. We can only gain the approximations of these true values. In this paper, by applying the K-class MPSVM classifier to a validation set, we can obtain the classification accuracies of K PSVM classifiers, a_i, $i = 1, \cdots, K$, $a_i \in [0,1]$. Let a validation set size be V, the *ith* PSVM correctly discriminates v_i examples in the validation set. The classification accuracy of the *i*th PSVM is

$$a_i = \frac{v_i}{V} \tag{27}$$

3) Design the bpa functions for every PSVM classifier. If the classification accuracy of a PSVM classifier satisfies $a < 1$, an unknown example will be correctly classified with probability a, that is, we can not determine which class it belong to with probability $(1-a)$. Thus, it is reasonable to set $m(\Theta) = 1 - a$. Suppose the probability output, decision output and classification accuracy of the *i*th PSVM classifier are p_i, d_i and a_i, respectively, the *bpa* functions are given by

$$m_i(\{k\}) = p_i a_i \tag{28}$$

$$m_i(\{1, 2, \cdots, k-1, k+1, \cdots, K\}) = (1 - p_i) a_i \tag{29}$$

$$m_i(\Theta) = 1 - a_i \tag{30}$$

where $i = 1, 2, \cdots, K$ and $k = i \in \Theta$. Obviously, formula (28)-(30) satisfies the condition (20).

4) Apply the Dempster-Shafer theory to combine evidences from all the individual K PSVM classifiers. All the combined values of mass, belief, plausibility, belief interval and ignorance can be obtained by using corresponding formula given in Section 3. There exist several criterions for giving final decision. In this paper, the maximal belief rule is used. Because all the non-singletons in focal elements just give meaningless decisions, only the singletons are compared with each other (A singleton is a subset with only one element).

The final decision is the class that corresponds to the singleton with the highest belief. Therefore, we can obtain the final decision function as follows.

$$d(x) = \arg \max_{i\in\{1,\cdots,K\}} \{bel(\{1\}), bel(\{2\}), \cdots, bel(\{K\})\} \quad (31)$$

The aforementioned procedure is illustrated in Fig. 1. A simple example is also given in the following to present the usage of DSMSVM.

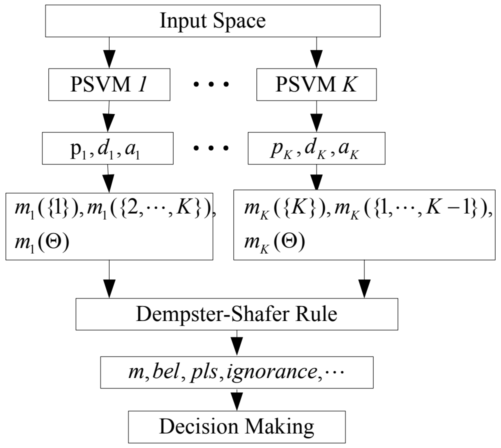

Fig.1. The method of Dempster-Shafer theory based multi-class support vector machines

Example: Given a 3-class classification problem, three PSVMs for binary classification are constructed. The accuracies of them are obtained by using a validation set (Table 1). For an unknown example, the individual probability outputs of these PSVMs are shown in Table 1. Our task is to determine which class the example belongs to.

Table 1 The data of an example

Individual classifier	Probability output	Accuracy
PSVM 1	0.37	0.95
PSVM 2	0.34	0.80
PSVM 3	0.38	0.90

According to (17), the decisions of three PSVMs are $\{2,3\}$, $\{1,3\}$ and $\{1,2\}$, that is to say, none of them can give a definitive decision. Obviously, by using MPSVM method (Expression (19)), we can determine the example belong to Class 3.

Now we will use DSMSVM to solve this problem. Let $\Theta = \{1,2,3\}$ be the frame of discernment, according to (28)-(30), the *bpa* are listed in Table 2. Then based on the formulas given in Section 3, the combination results of three PSVMs using Dempster-Shafer theory are given in Table 3.

Only the beliefs corresponding to singletons, $\{1\},\{2\},\{3\}$, is considered in Table 3. We can find the singleton $\{3\}$ has maximal belief value. Thus, the final decision is that the example belongs to Class 3 according to formula (31). We can also find the singleton $\{2\}$ with maximal plausibility and ignorance, which is mainly because the lowest accuracy of PSVM 2 results in large classification uncertainty. As compared with MPSVM and standard MSVM, DSMSVM provides a great deal of useful information for post-processing.

Table 2 The values of basic probability assignment (Groups: Gs; Values, Vs)

PSVM 1		PSVM 2		PSVM 3	
Gs	Vs	Gs	Vs	Gs	Vs
{1}	0.3476	{2}	0.2750	{3}	0.3332
{2,3}	0.6024	{1,3}	0.5250	{1,2}	0.5668
Θ	0.0500	Θ	0.2000	Θ	0.1000

Table 3 The combination results of three PSVMs (Groups:Gs; Mass: M; Belief: B; Plausibility: P; Ignorance: I)

Gs	M	B	P	I
{1}	0.3146	0.3146	0.3306	0.0160
{2}	0.3232	0.3232	0.3554	0.0322
{3}	0.3255	0.3255	0.3524	0.0270
{2,3}	0.0207	0.6694	0.6854	0.0160
{1,3}	0.0045	0.6446	0.6768	0.0322
{1,2}	0.0097	0.6476	0.6745	0.0270
Θ	0.0017	1.0000	1.0000	0

5. EXPERIMENTAL RESULTS

Tay and Shen (2003) proposed a method that rough sets theory is used to diagnose the valve fault for a multi-cylinder diesel engine. Due to the complex structure and multi-excite sources that exist in diesel engine, the vibration signals collated from the engine surface have the following characteristics. Four states are researched by Tay and Shen (2003): 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 were 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. The extracted single feature in frequency domain and time domain cannot indicate conspicuous difference exists among the different fault types. Therefore, 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).

Table 4 Classification accuracy of each part (training data: TND; testing data: TD)

Data set	1st part: TND 2nd part: TD	2nd part: TND 1st part: TD
Accuracy	0.78947	0.73684

Tay and Shen (2003) applied two-fold cross-validation test for showing the effect of the rough set theory in fault diagnosis. The classification accuracy is listed in Table 4. We can find that the average classification accuracy is 0.7632.

The aforementioned whole dataset is listed in (Shen, *et al.*, 2000). It consists of 37 instances, among which 25 instances are used as training set and the rest are used as test set in our experiment. The choice of kernel and of the regularizing parameter was determined via performance on a validation set. 80% of the training set is used for training binary SVM classifiers and the rest 20% of the training set is used as validation set. Three types of SVM for multi-class classification are used. The whole experiment is repeated for 50 times. In Table 5, the average classification accuracies are listed. We can find that all types of MSVMs have comparable performance. The classification accuracy obtained by using these MSVM methods largely outperforms that obtained by using Rough Set Theory (Table 4). However, the DSMSVM can provide more classification information than the standard MSVM and MPSVM.

Table 5 Classification accuracy of different MSVMs

Methods	Standard MSVM	MPSVM	DSMSVM
Accuracy (%)	93.83	93.33	93.50

6. CONCLUSIONS

Optional SVM based methods for multi-class classification problem are proposed in this paper.

The PSVM proposed by Platt (1999) can produce a posterior probability. Based on PSVM, the MPSVM using one-against-all strategy is proposed to deal with multi-class problem. The final decision is determined by the maximal probability rule.

It is almost impossible to obtain an ideal classifier to correctly discriminate all the unknown instances in inseparable cases. The error rate of classification can be considered as uncertainty lying in the classifier. Therefore, it is reasonable to represent MPSVM method in the frame of Dempster-Shafer theory. In this paper, the basic probability assignment is designed based on all the probability outputs and the performances of all PSVMs consisting in MPSVM. By using the Dempster-Shafer theory to combine all the evidences provided by every PSVM in MPSVM, and then applying the maximal belief rule only to the classes corresponding to the singletons, we can obtain the final decision output. In addition to this, the values of belief, plausibility, belief interval and ignorance about all the classes and class subsets are also gained. We call this new method as the Dempster-Shafer theory based multi-class support vector machine (DSMSVM). Compared with the standard MSVM and MPSVM, the DSMSVM provides bulk of useful information for post process. This makes it appropriate to be applied in some other fields, such as information fusion, including using Dempster-Shafer theory to combine multiple DSMSVMs, or combine them with other methods, for fewer weaker conditions need be satisfied for the application of Dempster-Shafer theory. In the experiment, the DSMSVM is applied to fault diagnosis for a diesel engine. The results show that our proposed method obtains a comparable performance with the standard MSVM and MPSVM. This provides a new optional method for multi-class classification problem and extends the application filed of SVM based method.

ACKNOWLEDGMENTS

This paper is supported by the National Key Fundamental Research Program of China (2002cb312200), and partially supported by the National High Technology Research and Development Program of China (2002AA412010) and the Natural Science Foundation of China (60174038).

REFERENCES

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Hsu, C.-W., and C.-J. Lin (2002). A Comparison of Methods for Multi-class Support Vector Machines. *IEEE Transactions on Neural Networks*, **13(2)**, 415-425.

Platt, J.C., N. Cristianini, and J. Shawe-Taylor (2000). *Large Margin DAGs for Multiclass Classification*. (Solla, S.A., T.K., Leen, and K.-R Müller (ED.)), 547–553. MIT Press:

Platt, J.C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. (Smola, A.J., P. Bartlett, B. Scholkopf, and D. Schuur-mans (ED.)), MIT Press.

Rakar, A., and P. BalleÂ (1999). Transferable Belief Model in Fault Diagnosis. *Engineering Applications of Artificial Intelligence*, **12**, 555-567.

Burges, C.J.C. (1998). *A Tutorial on Support Vector Machines for Pattern Recognition*. Kluwer Academic Publisher, Boston, Manufactured in The Netherlands.

Gill, P.E., W. Murray, and M.H. Wright (1981). *Practical Optimization*. Academic Press.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley.

Beynona, M., D. Coskerb, and D. Marshallb (2001). An Expert System for Multi-criteria Decision Making using Dempster Shafer theory. *Expert Systems with Applications*, **20**, 357-367.

Guan, J.W., and D.A. Bell (1992). *Evidence Theory and its Applications (Vol.1)*. North-Holland-Amsterdam, New York.

Tay, F.E.H., and L. Shen (2003). Fault Diagnosis based on Rough Set Theory. *Engineering Application of Artificial Intelligence*, **16**, 39-43.

Shen, L., F.E.H. Tay, L. Qu, and Y. Shen (2000). Fault Diagnosis using Rough Sets Theory. *Computers in Industry*, **43**, 61-72.