2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

Method of Combining Multi-class SVMs Using Dempster-Shafer Theory and its Application

Zhonghui Hu, Yuangui Li, Yunze Cai, and Xiaoming Xu

Abstract—**To deal with multi-source multi-class problems, the method of combining multiple multi-class probability support vector machines (MPSVMs) using Dempster-Shafer evidence theory is proposed. The MPSVM is designed by mapping the outputs of standard support vector machines into a calibrated posterior probability using a learned sigmoid function, and then combining these learned probability support vector machines. The Dempster-Shafer evidence theory is used to combine these learned MPSVMs. Two schemes of combination are composed. One of the schemes takes into account the prior information. Our proposed method is applied to the fault diagnosis of a diesel engine. The experimental results show that the accuracy and robustness of fault diagnosis can be improved significantly.**

I. INTRODUCTION

It is increasingly important to increase reliability and to decrease the possible loss of production due to unscheduled downtimes for machinery [1]. There exists intensive demand to understand what, where and how faults occur. Shen et al. [2] proposed a rough set theory based method that can diagnose more than one category of faults. One disadvantage 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 while continuous attributes exist. Because a prior knowledge about the attribute is hard to obtain, it is difficult to select an appropriate discretization method.

Support vector machines (SVMs) have been widely used in many fields. In most cases, the generalization performance of SVMs either matches or is better than that of competing methods [3]. SVMs are originally developed to solve binary classification problems. However, real world problems often require the discrimination between more than two categories. There are two approaches for constructing multi-class SVM (MSVM) [4]. One is by constructing and combining several standard SVM classifiers while the other is by directly considering all data in one optimization formulation. Hsu and Lin [4] indicated that the first approach is more suitable for practical use.

In general, a posterior probability produced by a classifier is convenient for post-processing. A sigmoid is trained to map the SVM outputs to the posterior probabilities in [5], while the sparseness of SVMs is still maintained. This method is called as the probability SVM (PSVM) method. We extend the one-against-all MSVM to the multi-class PSVM (MPSVM) in this paper.

In general, a posterior probability produced by a classifier is convenient for post-processing. A sigmoid is trained to map the SVM outputs to the posterior probabilities in [5], while the sparseness of SVMs is still maintained. This method is called as the probability SVM (PSVM) method. We extend the one-against-all MSVM to the multi-class PSVM (MPSVM) in this paper.

In data fusion area multi-source classification is an important research issue. The different or same types of information from several data sources are used for classification in order to achieve higher classification accuracy and robustness [6, 7]. Benediktsson et al. [7] points out conventional statistical pattern recognition methods are not appropriate in classification of multi-source data since such data cannot, in most cases, be modeled by a convenient multivariate statistical model. The neural-network models are superior to the statistical methods in terms of overall classification accuracy of training data. Furthermore, in many classification problems the SVM classifiers outperform the neural-network classifiers [3, 8]. However, for multi-source multi-class classification problem, the application of MSVMs is still an ongoing research area. One of the reasons is that its output is not a posterior probability. Therefore, the proposed MPSVM is promising to overcome this problem.

The probability outputs of MPSVMs proposed make the MPSVMs can be combined by using Dempster-Shafer

Manuscript received September 12, 2004. This work was supported by the National Key Fundamental Research Program of China (2002cb312200), the National High Technology Research and Development Program of China (2002AA412010), and in part supported by 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).

Yuangui Li is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: liyuangui@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).

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

0-7803-9098-9/05/$25.00 ©2005 AACC

evidence theory, a general extension of Bayesian theory that can robustly deal with incomplete data [9]. The evidence theory offers a number of advantages, including the opportunity to assign measures of probability to focal elements, and allowing for the attachment of probability to the frame of discernment. Furthermore, when the prior information is available, it can be taken into account in the extended Dempster-Shafer rule to improve the performance.

This paper is organized as follows. In Section II, the standard SVM for binary classification is reviewed. The PSVM method is also introduced. The one-against-all MPSVM is proposed in Subsection *C* of Section II. In Section III, the Dempster-Shafer evidence theory is introduced. In Section IV, two schemes of combining MPSVMs using the evidence theory are discussed. In Section V, our proposed method is applied to fault diagnosis for a diesel engine, and the experimental results are given. Finally, the conclusions are provided in Section VI.

II. Support Vector Machines

In this section, we first introduce the standard SVM for binary classification. Then, the PSVM is introduced. Finally, based on the PSVM, we develop a MPSVM algorithm based on the one-against-all strategy.

A. Support Vector Machines for binary classification

Given a training data set of (x_i, y_i) ($i = 1, \cdots, l$) where $x_i \in R^n$ and $y_i \in \{1, -1\}$ [10]. SVMs optimize the classification boundary by separating the data with the maximal margin hyperplane. The optimization problems in linearly inseparable and nonlinearly inseparable cases are discussed in the following [8].

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

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

The optimization problem (1) can be rewritten as

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

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

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

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

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

For the nonlinearly inseparable case, the original data are projected into a certain high dimensional Euclidean space H by a nonlinear map $\Phi : R^n \to H$. Introducing the kernel function $K\left(x_i, x_j\right)$ makes it not necessary to explicitly know $\Phi(\bullet)$ [3]. Hence, the more general kernel version of the optimization problem (1) is

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

The problem (5) can be rewritten as

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

The optimal hyperplane with maximal margin is obtained by solving (6)

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

The decision function that separates training data into two classes in the input space is

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

B. Probability Support Vector Machines

Platt [5] provided a kernel method which fits a sigmoid that maps SVM outputs to posterior probabilities, while still retains the sparseness of the SVM.

Suppose the continuous output of a standard SVM, the distance from the unknown instance x to the optimal classification hyperplane, be f. The parameters A and B of the parametric model $P(y = 1 \mid f)$ are adapted to give the best probability outputs.

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

As long as $A < 0$, the monotonicity of (9) is assured.

The parameters A and B can be found by minimizing the 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

$$p_i = P(y = 1 \mid f_i) \tag{11}$$

and

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

The optimization problem (11) is solved by using a model-trust minimization algorithm for robustness [11].

The model of out-of-sample data can prevent overfitting in training sigmoid. A simple out-of-sample model is used in [5]. 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. The probability of correct label is derived using Bayesian rule. Suppose N_+ positive examples are observed. The maximum a posteriori probability (MAP) 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

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

The non-binary target t'_i will converge to {0,1} when the training set size approaches infinity, which recovers the maximum likelihood sigmoid fit.

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

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

The corresponding decision function is

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

C. Multi-class Probability Support Vector Machines

Given a training data set $\{(x_i, y_i)\}_{i=1}^{l}$, where $x_i \in R^n$ and $y_i \in \{1, \cdots, K\}$. The objective is to correctly discriminate these classes from each other. Based on the PSVM and one-against-all strategy, the MPSVM can be constructed by applying the following procedure [12].

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

2) Construct *K* binary PSVM classifiers with an output $p_k(x)$, $k = 1, \cdots, K$ by training the corresponding sigmoid using the modified training set (f_i, t'_i), $i = 1, \cdots, l$.

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

$$d(x) = \arg\max\{p_1(x), \cdots\cdots, p_n(x)\} \tag{18}$$

III. THE DEMPSTER-SHAFER EVIDENCE THEORY

Dempster-Shafer evidence theory is regarded as a generalization of classic Bayesian theory. It explicitly represents ignorance as well as uncertainty, and offers a number of advantages, such as providing the opportunity to assign measures of probability to focal elements, and allowing for the attachment of probability to the frame of discernment [13, 14].

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

$$\begin{cases} m(\varnothing) = 0,\ (\varnothing \text{ - empty set}) \\ \sum_{x \in 2^\Theta} m(x) = 1 \end{cases} \tag{19}$$

where the notation 2^Θ is the power set of Θ. Any subset x of Θ with non-zero mass value is called a focal element.

A belief function $bel : 2^\Theta \to [0, 1]$ is defined by

$$bel(A) = \sum_{B \subseteq A} m(B), \text{ for all } A \subseteq \Theta. \tag{20}$$

It represents the confidence that a proposition lies in A or any subset of A.

The combined *bpa* function $m_1 \oplus m_2 \oplus \cdots \oplus m_N$: $2^\Theta \to [0, 1]$, which is used to combine the measures of evidence from different *N* sources, is defined by

$$(m_1 \oplus m_2 \oplus \cdots \oplus m_N)(A) = \frac{1}{K_N} \sum_{X_1 \cap X_2 \cap \cdots \cap X_N = A} m_1(X_1) m_2(X_2) \cdots m_N(X_N) \tag{21}$$

where K_N is defined as

$$K_N = \sum_{X_1 \cap X_2 \cap \cdots \cap X_N \neq \varnothing} m_1(X_1) m_2(X_2) \cdots m_N(X_N) \tag{22}$$

The normalization constant $1/K_N$ measures the extent of conflict between the pieces of evidence.

Sometimes the prior probabilities of hypotheses are helpful for improving the classification accuracy. Yen [15] showed that the subset-valued mapping used in the Dempster-Shafer theory can be extended to a probabilistic mapping to express the uncertainties of the evidence-hypotheses associations. In order to combine the hypothesis strength mass distributions $\mu_1, \mu_2, \cdots, \mu_n$ and prior probability P, Yen defined a respective mass function $c_1, c_2, \cdots, c_n$ called basic certainty assignment (*bca*) to discount the prior probabilities from each hypothesis strength mass distribution $\mu_1, \mu_2, \cdots, \mu_n$. The *bca* function is defined as

1) $c_i(\phi)=0$.

2) For every $A\subseteq\Theta$, $A\neq\phi$,

$$c_i(A)=\frac{\mu_i(\mathrm{A})/P(A)}{\sum_{\theta\subseteq\Theta,\theta\neq\phi}\left[\mu_i(\theta)/P(\theta)\right]}\text{, for } i=1,2,\cdots,n \qquad (23)$$

Hence, combination of hypothesis strength and prior probability is performed by the following steps.

Step 1: Transform *bpa* mass functions $\mu_1,\mu_2,\cdots,\mu_n$ from independent sources of evidence into respective *bca* mass functions $c_1,c_2,\cdots,c_n$ using the equation (23).

Step 2: All *bca* functions $c_1,c_2,\cdots,c_n$ are combined to $c_1\oplus c_2\oplus\cdots\oplus c_n$ using Dempster-Shafer's rule.

Step 3: The final combined bca function $c_1\oplus c_2\oplus\cdots\oplus c_n$ is transformed to a combined bpa mass function $\mu=\mu_1\oplus\mu_2\oplus\cdots\oplus\mu_n$ using the equations as below.

1) $(\mu_1\oplus\mu_2\oplus\cdots\oplus\mu_n)(\varnothing)=0$;

2) For every $A\subseteq\Theta$, $A\neq\varnothing$,

$$(\mu_1\oplus\mu_2\oplus\cdots\oplus\mu_n)(A)$$
$$=\frac{(c_1\oplus c_2\oplus\cdots\oplus c_n)(A)P(A)}{\sum_{\theta\subseteq\Theta,\theta\neq\varnothing}[(c_1\oplus c_2\oplus\cdots\oplus c_n)(\theta)P(\theta)]} \qquad (24)$$

IV. Combination of Multiple MPSVMs

In this section, two schemes of combining MPSVM classifiers using Dempster-Shafer theory are proposed. The prior information is not taken into account in the scheme I (Fig. 1), whereas it is considered in Scheme II (Fig. 2).

Suppose there are *J* data sources and exist exhaustive *K* patterns (hypotheses) in these data sources. *J* MPSVM classifiers can be constructed. Set the probability output of the *k*th PSVM in the *j*th MPSVM is $p_{jk}(x)$, $k\in[1,K]$, $j\in[1,J]$. The probability $p_{jk}(x)$ should be normalized for combination by using the following equation

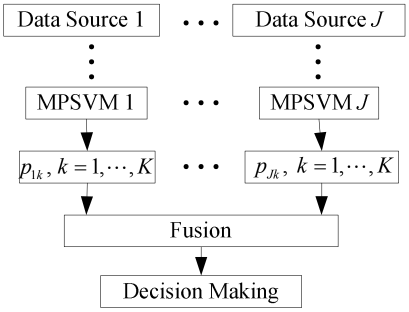

Fig. 1. Combination of MPSVMs without prior information

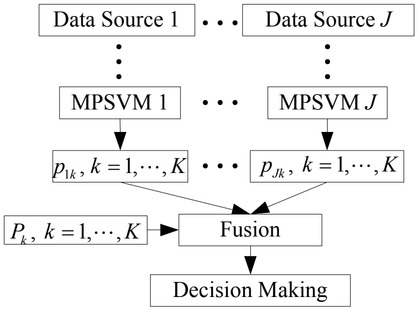

Fig. 2. Combination of MPSVMs considering prior information

$$\overline{p}_{jk}(x)=\frac{p_{jk}(x)}{\sum_{k=1,\cdots,K}p_{jk}(x)} \qquad (25)$$

The *bpa* function m_{jk} is defined as

$$m_{jk}(\{k\}_x)=\overline{p}_{jk}(x)\text{, } k=1,\cdots,K\text{ , } j=1,\cdots,J \qquad (26)$$

The combination is performed as below.

For an unknown instance x , the normalization constant is

$$K_J(x)=\sum_{k=1,\cdots,K}\overline{p}_{1k}(x)\overline{p}_{2k}(x)\cdots\overline{p}_{Jk}(x) \qquad (27)$$

Therefore, the combined mass function of *K* MPSVM classifiers is defined as

$$m_k(\{k\}_x)=\frac{1}{K_J(x)}\overline{p}_{1k}(x)\overline{p}_{2k}(x)\cdots\overline{p}_{Jk}(x)\text{, } k=1,\cdots,K \qquad (28)$$

Notice the fact that all focal elements of mass function *m* are singletons, we can obtain

$$bel_k(\{k\}_x)=m_k(\{k\}_x)\quad\text{, } k=1,\cdots,K \qquad (29)$$

The decision function based on the maximal belief rule is

$$d(x)=\arg\max_{\{1,\cdots,K\}}\{bel_1(\{1\}_x),\cdots,bel_K(\{K\}_x)\} \qquad (30)$$

For Scheme II, the prior probabilities hidden in the data set is determined by

$$P_{jk}=\frac{l_{jk}}{L_j}\text{, } j=1,\cdots,J\text{ , } k=1,\cdots,K \qquad (31)$$

where P_{jk} is the prior probability of the *k*th pattern, L_j is the size of the training data set and l_{jk} is the size of instances belonging to pattern *k*.

For an unknown instance x , the *bca* is first calculated by

$$c_{jk}(\{k\}_x)=\frac{\overline{p}_{jk}(x)/P_{jk}}{\sum_{k=1,\cdots,K}\left[\overline{p}_{jk}(x)/P_{jk}\right]} \qquad (32)$$

Secondly, we calculate the combined mass functions

$$c_k(\{k\}_x)=\frac{1}{KC_J(x)}c_{1k}(x)c_{2k}(x)\cdots c_{Jk}(x) \qquad (33)$$

where $KC_J(x)=\sum_{k=1,\cdots,K}c_{1k}(x)c_{2k}(x)\cdots c_{Jk}(x)$.

Finally, we can obtain the combined *bpa* mass function by using the following equation

$$m_k(\{k\}_x)=\frac{(c_{1k}\oplus c_{2k}\oplus\cdots\oplus c_{JK})(\{k\}_x)P_k}{\sum_{\theta\subseteq\Theta,\theta\neq\varnothing,k=1,\cdots,K}(c_{1k}\oplus c_{2k}\oplus\cdots\oplus c_{JK})(\theta)P_k} \qquad (34)$$

The final decision function based on the strategy of maximal belief function is

$$d(x)=\arg\max_{\{1,\cdots,K\}}\{bel_1(\{1\}_x),\cdots,bel_K(\{K\}_x)\} \qquad (35)$$

where $bel_k(\{k\}_x)=m_k(\{k\}_x)$, $k=1,\cdots,K$.

V. EXPERIMENTAL RESULTS

The dataset given by Shen et al. [2] is used to diagnose the valve fault for a multi-cylinder diesel engine. The vibration signals are collated from three sampling point on the engine surface. Due to the complex structure and multi-excite sources that exist in diesel engine, it is difficult to analyze these data. Four states are researched in [2]: 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 at 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. Thus, three datasets from corresponding three sampling points (data sources) are constructed.

The dataset is divided into two parts equally in cross-validation test for showing the effect of the rough set theory in fault diagnosis [16]. The classification accuracy excerpted from [16] is listed in Table I.

The whole dataset consists of 37 instances, among which 25 instances are used as training set and the rest are used as testing set. The kernel and regularizing parameters are determined by using a validation set. Eighty percent of the training set is used for training SVM classifiers and the rest is used as a validation set. The whole experiment is repeated 50 times, where the training set and testing set are randomly selected without replacement every time. The experimental results are given in Table II. The accuracy of three MPSVM classifiers corresponding to three sampling points is given in D1, D2 and D3 columns. The average accuracy of proposed methods is listed in S.I and S.II. Table II shows the accuracy of fault diagnosis is improved by using the proposed methods. When taken into account the prior information hidden in the dataset, the best accuracy is obtained. We also find the robustness of fault diagnosis is also improved.

TABLE I
CLASSIFICATION ACCURACY OF EACH PART

Data set	1: training data 2: testing data	2: training data 1: testing data
Accuracy	0.78947	0.73684

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

Scheme	Single data source			S.I	S.II
	D1	D2	D3		
Average accuracy	90.83	82.00	91.50	95.83	97.33
Minimal accuracy	66.67	58.33	75.00	75.00	83.33
Standard deviation	8.95	11.83	5.45	7.58	5.94

VI. CONCLUSIONS

To deal with distributed multi-source multi-class problem, Demspter-Shafer evidence theory is used to combine multiple MPSVM classifiers corresponding to the data sets from different sources in this paper. The final decision is based on the maximal belief principle.

Two strategies are proposed. Only the outputs of multiple MPSVM classifiers are combined by using combination rule in Scheme I. Extend Dempster-Shafer rule is applied to Scheme II, which takes full advantages of the prior information hidden in data sets.

Our proposed methods are evaluated by applying to fault diagnosis of a diesel engine. The accuracy and robustness of fault diagnosis is improved.

REFERENCES

[1] B. S. Yang, T. Han and Y. S. Kim, "Integration of ART-Kohonen Neural Network and Cased-based Reasoning for Intelligent Fault," *Expert Systems with Applications*, Vol. 26, pp. 387-395, 2004.

[2] 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.

[3] 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.

[4] 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.

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

[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, pp54-64, 1997.

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

[9] Andrej Rakar and Peter BalleÂ, "Transferable Belief Model in Fault Diagnosis," *Engineering Applications of Artificial Intelligence*, Vol. 12, pp. 555-567, 1999.

[10] K. M. Lin and C. J. Lin, "A Study on Reduced Support Vector Machines," IEEE Trans. Neural Networks, Vol.14, no. 6, pp. 1449-1459, 2003.

[11] P.E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, Academic Press, 1981.

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

[13] M. Beynona, D. Coskerb and D. Marshallb, "An expert system for multi-criteria decision making using Dempster Shafer theory," *Expert Systems with Applications*, Vol. 20, pp. 357-367, 2001.

[14] J. W. Guan and D. A. Bell, *Evidence Theory and its applications*, Vol.1, North-Holland-Amsterdam, New York, 1992.

[15] J. Yen, "GERTIS: A Dempster-Shafer Approach to Diagnosing Hieraractical Hypotheses," *Communications of the ACM*, Vol. 5, no. 32, pp. 573-585, 1989.

[16] 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.