A Robust Fuzzy Support Vector Machine for Two-class Pattern Classification

G. H. Lee, J. S. Taur, and C.W. Tao

Abstract

This paper proposes a systematic method to classify data with outliers. The essential techniques consist of the outlier detection and the fuzzy support vector machine (FSVM). In this approach, the main body set for each class is first determined by the outlier detection algorithm (ODA) that estimates the outliers based on the total similarity objective function. Then, incorporated with the total similarity measure of the ODA, a fuzzy membership degree is assigned to each training sample. Experiments show that the proposed method can greatly reduce the effects of outliers in the training process and the final decision surface of the FSVM is insensitive to outliers.

Keywords: Outlier detection, Support vector machines, Fuzzy SVMs.

1. Introduction

The theory of support vector machines (SVMs) first developed by Vapnik and his research group is a powerful methodology for solving pattern classification and regression estimation problems [1], [2], [3], [4]. Those techniques are based on the theoretical learning theory that embodies the structural risk minimization (SRM) principle [2]. SVMs have been shown to provide high generalization performance on a wide range of applications. The SVM technique can be considered as an alternative training method for polynomial-function, radial-basis-function, and multilayer-perceptron classifiers by selecting proper kernel functions. In recent years, SVMs have been applied broadly and successfully to various fields such as pattern recognition [3], image classification [5], time prediction [6], and regression [7], [8].

In the theory of SVM, one of the main assumptions is that all data in the training set are treated equally. However, noisy data or outliers are usually inevitable in practical applications [9]. This may make the decision surface deviate severely from the optimal hyperplane due to the unawareness of outliers in the training process and lead to the degradation of generalization performance in the test stage. Recently, some algorithms have been proposed to tackle the outlier problem. In [10], the weighted least square support vector machine (LS-SVM) is proposed to reduce the effects of outliers. Nevertheless, the parameters in those techniques need to be chosen carefully. In [11], an adaptive margin SVM is proposed based on the utilization of adaptive margins for each training pattern. However, there is no general way to use the class center in the margin of each training data to suppress the effects of noise and outliers. Using of the distance between each data point and the center of the respective class, a robust SVM [5] is proposed to calculate the adaptive margin which makes the SVM less sensitive to outliers. In [12], a robust SVM based on an accelerated decomposition algorithm is proposed to solve the over-fitting problem that results from outliers in the training data set. This approach and the technique in [5] both depend on the distance between the training data and class centers in the feature space. In addition, some fuzzy support vector machines (FSVMs) have been proposed to tackle the outlier problem [13], [14]. Huang et al., [13] adopt a fuzzy *c*-means algorithm cascaded with an unsupervised neural network to detect outliers in a training data set. Then, a membership model is developed to assign membership values to main body training samples and outliers according to their relative importance in the training data set. Therefore, FSVMs can reduce the over-fitting effects and outperform SVMs in classification problems with outliers.

In this paper, a systematic method using the outlier detection algorithm (ODA) and the FSVM is proposed to handle the outlier problem. As shown in Figure 1, the proposed approach first adopts the ODA to obtain the main body from all the training samples in each class. Incorporated with the total similarity measure from the ODA, each sample can be properly assigned a membership degree by a pre-selected sigmoid function. Thus the proposed approach enhances the insensitivity of the FSVM to outliers.

The rest of this paper is organized as follows. Section 2 gives a brief review of SVMs and FSVMs. In Section

Corresponding Author: C. W. Tao is with the Department of Electrical Engineering, National ILan University.
E-mail: cwtao@niu.edu.tw
Manuscript received 28 Aug. 2003; revised 1 Dec. 2003; accepted 21 June. 2004.

3, the proposed algorithm is developed. The experimental results and discussions are presented in Section 4. Finally the conclusion is given in Section 5.

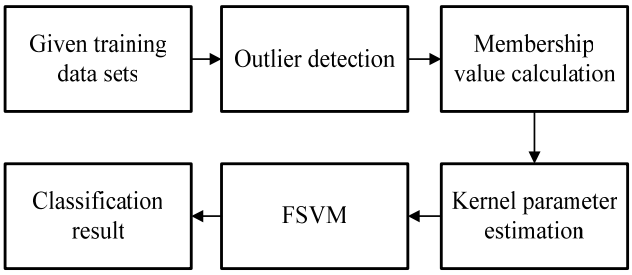


Figure 1. The proposed classification system.

2. Related Background

In this section, we will briefly review the algorithms of SVMs and FSVMs. More detailed descriptions can be found in [1], [2], [13], and [14].

A. *Support Vector Machines*

The support vector machine is a classifier based on the structural risk minimization to find the hyperplane that maximizes the margin between classes. Without loss of generality, the theory of SVMs is introduced through a two-class classification problem. Assume that the samples from class one and class two are associated with a class label $y_i = -1$ and $y_i = +1$ respectively. Given a training data set S of N data points

$$S = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \tag{1}$$

in which the i^{th} input sample $\mathbf{x}_i \in \mathbb{R}^n$ belongs to one of the two classes labeled by $y_i \in \{+1, -1\}$. The training goal of the SVM is to find an optimal hyperplane $\mathbf{w}^T\varphi(\mathbf{x}) + b = 0$ that maximally separates the two classes of training samples, where $\varphi(\bullet)$ is a nonlinear function which maps the input space into a higher dimensional space, $\mathbf{w}$ is a weight vector, and $b \in \mathbb{R}$ is a bias of the hyperplane. Then the sample point $\mathbf{x}_i$ can be assigned its corresponding class label and the classifier can be expressed as

$$g(\mathbf{x}_i) = \text{sgn}\left(\mathbf{w}^T\varphi(\mathbf{x}_i) + b\right) \tag{2}$$

where $\text{sgn}(\bullet)$ stands for the bipolar sign function. For a separable case, there exists a weight vector $\mathbf{w}$ and a bias b such that each sample point satisfies the following conditions:

$$\begin{cases} \mathbf{w}^T\varphi(\mathbf{x}_i) + b > +1, & for \quad y_i = +1 \\ \mathbf{w}^T\varphi(\mathbf{x}_i) + b < -1, & for \quad y_i = -1 \end{cases} \tag{3}$$

which are equivalent to

$$y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq +1, \quad i = 1, 2, ..., N. \tag{4}$$

In this separable case, the optimal hyperplane that maximizes the margin of separation can be found. However, in the non-separable case, the separating hyperplane in the higher dimensional space does not exist. In order to handle such cases, a set of nonnegative slack variables $\{\xi_i\}_{i=1}^{N}$ is introduced such that the following conditions are satisfied

$$y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, ..., N \tag{5}$$

$$\xi_i \geq 0, \quad i = 1, 2, ..., N. \tag{6}$$

This approach allows training samples that violate Eq. (4). According to the structural risk minimization, the optimal decision can be found by solving the following quadratic programming (QP) problem:

$$\min \quad \mathbb{L}(\mathbf{w}, \xi_i) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{7}$$

subject to Eqs. (5) and (6), where C is a predefined positive constant. A smaller C imposes a less penalty on empirical errors. Instead of solving the QP optimization in the primal space, a set of Lagrange multipliers is introduced for Eqs. (5) and (6), and the following dual problem can be obtained

$$\max_{\alpha_i} \; Q(\alpha_i) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j) \tag{8}$$

$$subject\ to \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, ..., N \tag{9}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0, \quad i = 1, 2, ..., N. \tag{10}$$

According to the Kuhn-Tucker conditions, the solution $\{\alpha_i\}_{i=1}^{N}$ to Eqs. (8), (9), and (10) has to satisfy the following conditions

$$\alpha_i\left(y_i(\mathbf{w}_0^T\varphi(\mathbf{x}_i) + b_0) - 1 + \xi_i\right) = 0, \quad i = 1, 2, ..., N \tag{11}$$

$$\left(C - \alpha_i\right)\xi_i = 0, \quad i = 1, 2, ..., N. \tag{12}$$

Those points with $\alpha_i > 0$ are called support vectors which can be divided into two types. If $0 < \alpha_i < C$, the corresponding training points just lie on one of the margins. If $\alpha_i = C$, this type of support vectors are regarded as misclassified data.

B. *Fuzzy Support Vector Machines*

The SVM has been introduced as a powerful tool for solving classification problems. However, there are still some difficulties in applying the theory to practical problems. One of the major difficulties is that the SVM algorithm is sensitive to outliers. Although the influence of outliers can be reduced by choosing a proper parameter *C*, it is not easy to find a suitable *C*. In the formulation in Eq. (7), the parameter *C* is a user-defined parameter to penalize the training data with a positive ξ_i. A larger *C* imposes a heavier penalty on the misclassified training data and thus results in fewer support vectors and a narrower separation region, while a smaller *C* sets a smaller penalty for the error and thus leads to a wider margin [16]. It is clearly that all training points in the class are treated equally in the theory of the SVM. This may make the SVM very sensitive to noise and outliers [9]. In many applications of pattern classification, some training points are more important than the others. Therefore, it is very important to distinguish the meaningful training points from outliers or noisy samples. This can be achieved by assigning a membership value u_i to each training point x_i.

Assume that a training data set S_{Fuzzy} of *N* data points with corresponding membership values is given by

$$S_{Fuzzy} = \{\mathbf{x}_i, y_i, u_i\}_{i=1}^{N} \tag{13}$$

where $\mathbf{x}_i \in \mathbb{R}^n$ is the i^{th} input data sample, $y_i \in \{-1,+1\}$ is its label, and $0 \le u_i \le 1$ is its membership value. In contrast with SVMs, the term $u_i\xi_i$ is used as a weighted measure of the error in FSVMs. Because a proper membership value is assigned to each training sample, FSVMs should be more robust in the classification problems. In this formulation, the optimal separating hyperplane is regarded as the solution to

$$\min \quad \mathbb{L}(\mathbf{w},\xi_i) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} u_i\xi_i \tag{14}$$

$$\textit{subject to} \quad y_i(\mathbf{w}^T\varphi(\mathbf{x}_i)+b) \ge 1-\xi_i, \quad i=1,2,...,N \tag{15}$$

$$\xi_i \ge 0, \quad i=1,2,...,N \tag{16}$$

where *C* is a positive parameter to be defined by the user. The larger (smaller) the value u_i is, the more (less) influence the parameter ξ_i has, and the more (less) important the training point $\mathbf{x}_i$ is. Now the Lagrangian function can be constructed as

$$\begin{aligned} Q(\mathbf{w},b,\xi_i,\alpha_i,v_i,u_i) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} u_i\xi_i \\ &-\sum_{j=1}^{N}\alpha_i\left(y_i(\mathbf{w}^T\varphi(\mathbf{x}_i)+b)-1+\xi_i\right) - \sum_{i=1}^{N} v_i\xi_i \end{aligned} \tag{17}$$

where $\{\alpha_i, v_i\}_{i=1}^{N}$ are Lagrange multipliers. The solution is given by the saddle point of the Lagrangian function $Q(\mathbf{w},b,\xi_i,\alpha_i,v_i,u_i)$. Thus, the following dual optimization problem (function of α_i only) can be obtained

$$\max_{\alpha_i} \; Q(\alpha_i) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j) \tag{18}$$

$$\textit{subject to} \quad 0 \le \alpha_i \le u_i C, \quad i=1,2,...,N \tag{19}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0, \quad i=1,2,...,N \tag{20}$$

and the Kuhn-Tucker conditions are

$$\alpha_i\left(y_i(\mathbf{w}_0^T\varphi(\mathbf{x}_i)+b_0)-1+\xi_i\right) = 0, \quad i=1,2,...,N \tag{21}$$

$$\left(u_i C-\alpha_i\right)\xi_i = 0, \quad i=1,2,...,N. \tag{22}$$

When compared with SVMs, it is obvious that the upper bounds of FSVMs in Eq. (19) are different from those of SVMs. The upper bounds of FSVMs are function of membership values u_i such that they can adjust the weighting for meaningful points and outliers in the training process, while the upper bounds of SVMs are constants. Similarly, there are also two kinds of support vectors. Training points with $0 < \alpha_i < u_i C$ will lie on the margin of the hyperplane. Training points with $\alpha_i = u_i C$ are misclassified. Once the values of α_i have been found, the solution of $\mathbf{w}_0$ can be determined by

$$\mathbf{w}_0 = \sum_{i=1}^{N_{SV}}\alpha_i y_i \varphi(\mathbf{x}_i) \tag{23}$$

where N_{SV} is the number of support vectors, and the value of threshold b_0 can be obtained from Eq. (21). In this way, the decision function can be obtained in the new feature space, i.e.,

$$g(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i=1}^{N_{SV}}\alpha_i y_i \varphi(\mathbf{x}_i)\varphi(\mathbf{x}) + b_0\right]. \tag{24}$$

The calculation of Eqs. (8), (18), and (24) requires the computation of the inner product $\varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j)$ or $\varphi(\mathbf{x}_i)\varphi(\mathbf{x})$ in a high dimensional feature space. By using a suitable "Kernel function K", ($K(\mathbf{x}_i,\mathbf{x}_j)$

$= \varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j)$), that obeys the Mercer condition [1], [15], Eqs. (8) and (18) can be computed in the input space through this kernel technique. Similarly, the discriminant function can be written as

$$g(\mathbf{x}) = \text{sgn}\left[\sum_{i=1}^{N_{SV}} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b_0\right]. \quad (25)$$

The major advantage of using a kernel function is that the explicit computation of $\varphi(\mathbf{x}_i)$ can be avoided. Instead of calculating the inner product of $\varphi(\mathbf{x}_i)\varphi(\mathbf{x})$ in the feature space, the kernel function $K(\mathbf{x}_i,\mathbf{x})$ can be obtained in the primal input space.

3. The Proposed Techniques

In this section, a systematic method is developed to classify data sets containing outliers. It includes an ODA for detecting outliers, the FSVM machine with the fuzzy membership function, and the kernel parameter estimation.

A. *Outliers Detection*

In this work, we introduce a simple but effective method based on the similarity measure which aims to detect the outliers in a training set. It is assumed in this algorithm that outliers in a class have two important characteristics [17], [18]. The first one is that the number of outliers should be much smaller than the number of patterns in the main body. The second one is the outliers should at least somewhat separate from the main body. More precisely, let $S = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ be a set of training samples from a class distributed in an unknown probability density function in the input space. We use a similarity measure function (SMF) denoted as $\mathcal{D}(\mathbf{x}_i, S)$ to measure whether the data $\mathbf{x}_i$ is located inside, near or far from the main body of S. The SMF $\mathcal{D}(\mathbf{x}_i, S)$ is defined as

$$\mathcal{D}(\mathbf{x}_i, S) = \sum_{j=1}^{m}\left(\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right)\right)^{\lambda}, \quad \lambda > 0 \quad (26)$$

where

$$\beta = \frac{1}{m}\sum_{\mathbf{x}_j \in S}\|\mathbf{x}_j - \mathbf{z}\|^2, \quad \mathbf{z} = \frac{1}{m}\sum_{\mathbf{x}_j \in S}\mathbf{x}_j, \quad (27)$$

and $\|\bullet\|$ denotes the Euclidean distance. In Eq. (26), λ is an adjustable parameter that is used to replace the effect of the parameter β. Therefore, the parameter β can be assigned a fixed value, for example, the sample variance of the data. Eq. (26) is also used in the mountain clustering method proposed by Yager and Filev [19], [20]. With a proper parameter λ, it can be regarded as the estimate of the density shape of the data points in the neighborhood of $\mathbf{x}_i$. The data point $\mathbf{x}_i$ with a smaller $\mathcal{D}(\mathbf{x}_i, S)$ will be located farther away from the main body of S. When there are more data points around $\mathbf{x}_i$, the value $\mathcal{D}(\mathbf{x}_i, S)$ will become larger. Thus the set $\{\mathcal{D}(\mathbf{x}_i, S)\}_{i=1}^{m}$ can be treated as an index to detect the outliers.

In order to analyze the effect of the parameter λ, the similarity measure for a spiral-shaped data set S with 3 outliers is calculated using Eq. (26). Plots (b), (c), and (d) in Figure 2 show the results of the SMF with λ=1, 10, and 20, respectively, and β is pre-selected as the sample variance. From Figure 2 (b), it is clear that the SMF with λ=1 can not isolate the outliers from the main body of S. However, after increasing λ to 10 or 20, the peaks of the SMF can clearly distinguish the outliers from the main body. It should be noted that the peak values of the main body are much larger than those of the outliers. Although Figures 2 (b), (c), and (d) can provide the visual impression for selecting the value of the parameter λ, a method for selecting a proper λ must be taken into account for a systematic method.

In an application, the percentage of the outliers in the data set S is unknown. In order to detect all the outliers, a larger initial value δ% is adopted to represent the candidates for outliers. According to Figure 2 and Eq. (26), increasing λ is equivalent to decreasing the neighborhood radius of the data point $\mathbf{x}_i$. For a data point $\mathbf{x}_i$, Eq. (26) can be rewritten as

$$\mathcal{D}(\mathbf{x}_i, S) = 1 + \sum_{\substack{j=1 \\ j\neq i}}^{m}\left(\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right)\right)^{\lambda}, \quad \lambda > 0 \quad (28)$$

where the second term in the right side of the equality represents the contribution from the other elements in the data set S to $\mathcal{D}(\mathbf{x}_i, S)$. If the data point $\mathbf{x}_i$ is an outlier and λ has been properly determined, the value of the second term should be very small. Thus, we first select a proper λ such that $\mathcal{D}(\mathbf{x}_i, S) < \theta_T$ for the data points $\mathbf{x}_i$ with the δ% lowest value of the $\mathcal{D}(\mathbf{x}_i, S)$ in the data set. Then the number of outliers can be estimated. In our experiments, the initial threshold is set to be

$$\theta_T = 1 + 0.02 \times (m - 1) \quad (29)$$

and its maximal value is $\theta_T = 4$. The ODA for a given class can be summarized as following steps :

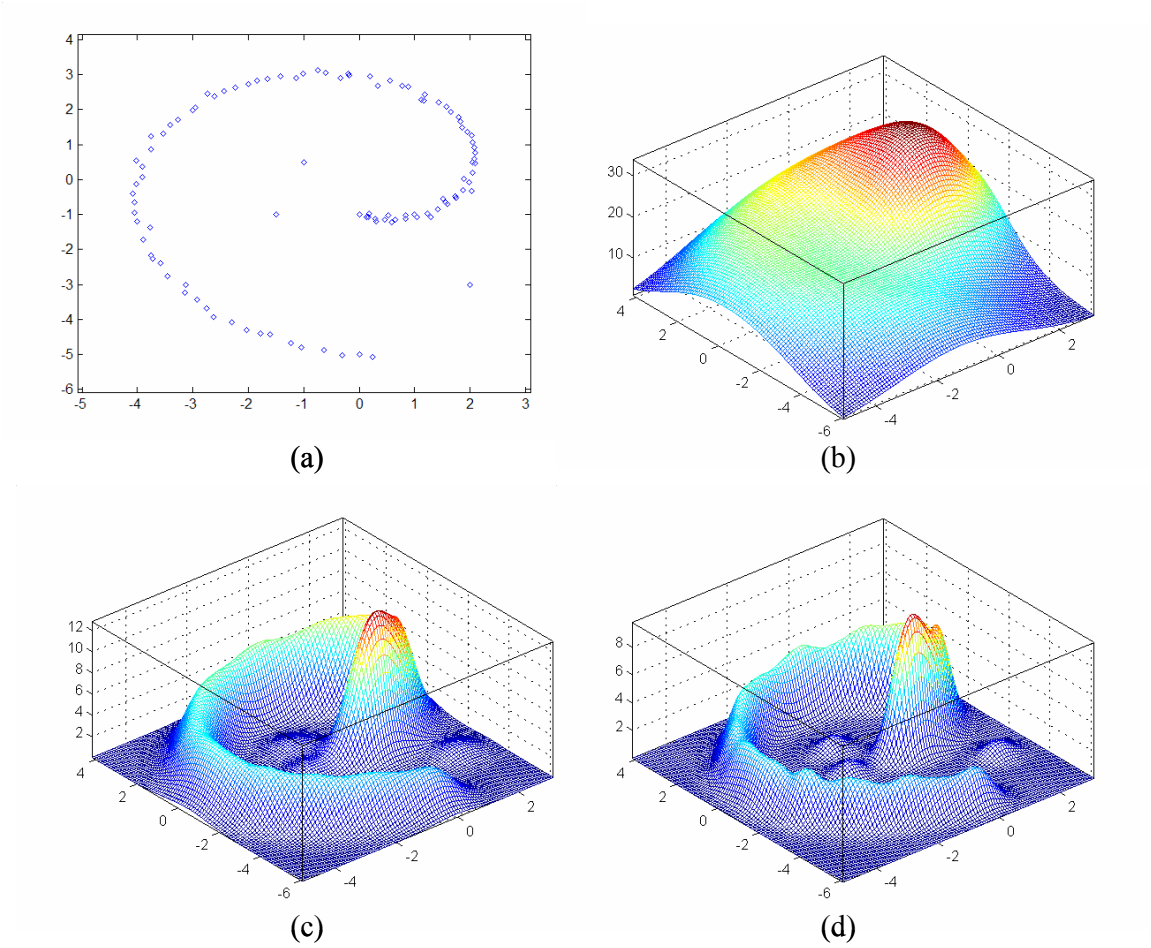


Figure 2. (a) A data set with spiral shape. (b), (c), and (d) are 3D plots using Eq. (26) with λ=1, 10, and 20, respectively.

Step 1) Set λ=1, $\Delta\lambda$=0.1, δ%=15%, β as a fixed value, $\gamma = \mathbf{int}(m \times \delta\%)$, and *outlier_loop* = **true**.

Step 2) Calculate $\{\mathcal{D}(\mathbf{x}_i, S)\}_{i=1}^{m}$ and rearrange the corresponding values in ascending order denoted as $\{g_i\}_{i=1}^{m}$.

Step 3) **If** $g_\gamma \le \theta_T$, **then**

GOTO **Step 4**

Else

$\lambda = \lambda + \Delta\lambda$

GOTO **Step 2**

End

Step 4) **If** (*outlier_loop*), **then**

$\mathbf{P}_{diff} = \{g_{i+1} - g_i\}_{i=1}^{\gamma-1}$

$P_{\max} = \max_i \mathbf{P}_{diff}$

$P_{mean} = \frac{1}{\gamma - 2}\sum\left(\mathbf{P}_{diff} - \{P_{\max}\}\right)$

If $(P_{max} - 3P_{mean}) > 0.1$, **then**

$\gamma_{outlier} = \arg\max_{i=1}^{\gamma-1} \mathbf{P}_{diff}$

Else

$\gamma_{outlier} = \arg_i \{g_i \ge (g_1 + 0.5(g_\gamma - g_1))\}$

End

$\gamma = \gamma_{outlier}$

$\theta_T = g_\gamma$

outlier_loop = **false**

$\lambda = 1$

GOTO **Step 2**

Else

$\lambda_e = \lambda$

$\tau_b = \frac{\beta}{\lambda_e}$

Stop the ODA

End

It should be noted that the pre-determined value β

must be large enough such that $g_\gamma > \theta_T$ for the initial value λ=1. $\gamma_{outlier}$ is the estimated number of outliers and λ_e is the estimate of λ. Finally, the main body of the data set S and the parameter $\tau_b = \beta / \lambda_e$ for the main body can be obtained. For the convenience of descriptions, the superscripts "+" and "−" are used to denote variables for Class 1 and Class 2, respectively. Thus, $S^+ = \{\mathbf{x}_i^+\}_{i=1}^{m^+}$ ($S^- = \{\mathbf{x}_i^-\}_{i=1}^{m^-}$) denotes the data set with an output label $y_i = +1$ ($y_i = -1$) in rest of the paper, where m^+ (m^-) is the number of samples in S^+ (S^-). In order to evaluate the ODA, a data set with two classes is shown in Figure 3 (a), where data points with the "◊" ("∗") sign represent S^+ (S^-) and each class contains 6 outliers. Figure 3 (b) shows the value of the SMF function for each data point with δ%=15%. It is obvious that the amplitudes of the outliers are much lower than those of the main body. After applying the ODA, the main body of S^+ (S^-) with the corresponding $\lambda_e^+ = 5.0$ ($\lambda_e^- = 4.7$) is shown in Figure 3 (c), the outliers for each class are identified correctly, and the bandwidth for the main body of S^+ (S^-) is $\tau_b^+ = 1.04$ ($\tau_b^- = 1.13$). Figure 3 (d) shows that the obtained λ_e's for the SMF can represent the density shape of the two main body sets.

B. *Membership Functions for FSVMs*

In general, the criterion to choose the membership value for each sample depends on the relative importance of the data point in its class. As described previously, the ODA is capable of fitting the density shape of a given data set, and the estimated $\tau_b = \beta / \lambda_e$ can represent the suitable interpolation neighborhood radius of a data point in the given data set. Thus, incorporated with the similarity measure obtained from the ODA, a membership value can be assigned to every training sample by the following fuzzy model to form the training fuzzy sets, cf. Eq. (13). The membership value set for S^+ and S^- is defined as

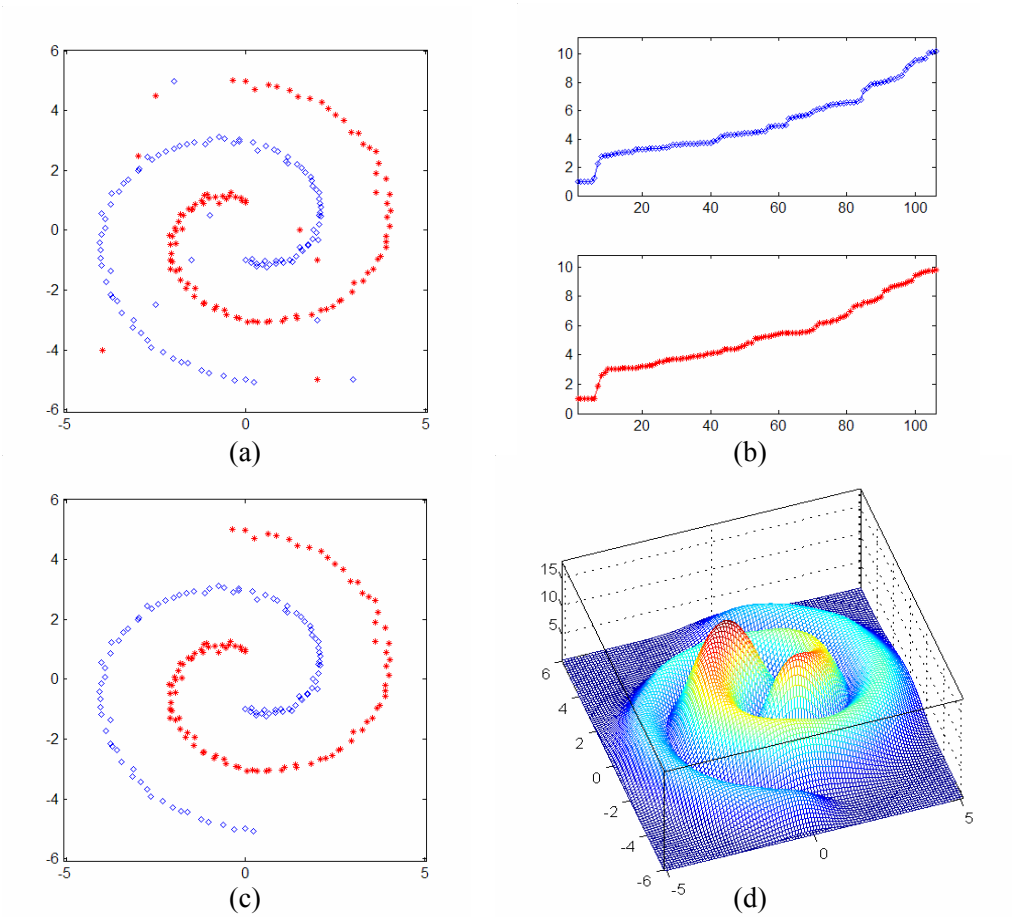


Figure 3. (a) A data set of two classes with outliers. (b) The amplitudes of the SMF function for each data point. (c) The corresponding main body set after applying the ODA technique. (d) The density shape of the main body sets with λ_e=5.0, and 4.7, respectively.

$$\mathbf{u} = \left\{u_1^+, ..., u_{m^+}^+, u_1^-, ..., u_{m^-}^-\right\} \tag{30}$$

where $u_i^+ = \min\{u_i^{++}, u_i^{+-}\}$ and $u_i^- = \min\{u_i^{--}, u_i^{-+}\}$. u_i^{++} is an index of the importance of the data point $\mathbf{x}_i^+$ in the data set S^+, i.e.,

$$u_i^{++} = \frac{1}{1+\exp\left(-\eta_1\left(\mathcal{D}(\mathbf{x}_i^+, S^+) - \eta_2\theta_T^+\right)\right)} \tag{31}$$

with

$$\mathcal{D}(\mathbf{x}_i^+, S^+) = \sum_{j=1}^{m^+}\left(\exp-\left(\frac{\left\|\mathbf{x}_i^+ - \mathbf{x}_j^+\right\|^2}{\beta^+}\right)\right)^{\lambda_e^+}, \quad \lambda_e^+ \geq 1 \tag{32}$$

where λ_e^+ is obtained from the ODA, β^+ is a pre-determined value, and η_1 controls the change rate of the sigmoid function. $\mathcal{D}(\mathbf{x}_i^+, S^+)$ measures the similarity of $\mathbf{x}_i^+$ in S^+, θ_T^+ is determined by the ODA algorithm, and η_2 is a weighted factor. In contrast to u_i^{++}, u_i^{+-} measures the overlapping degree between the data point x_i^+ and all elements in the data set S^- and is defined as

$$u_i^{+-} = 1 - \frac{1}{1+\exp\left(-\eta_3\left(\mathcal{D}(\mathbf{x}_i^+, S^-) - \eta_4\theta_T^{+-}\right)\right)} \tag{33}$$

with

$$\mathcal{D}(\mathbf{x}_i^+, S^-) = \sum_{j=1}^{m^-}\left(\exp-\left(\frac{\left\|\mathbf{x}_i^+ - \mathbf{x}_j^-\right\|^2}{\beta^-}\right)\right)^{\lambda_e^-}, \quad \lambda_e^- \geq 1 \tag{34}$$

where β^- and θ_T^{+-} are pre-determined value for S^-, η_3 control the change rate of the sigmoid function, η_4 is a weighted factor, and $\mathcal{D}(\mathbf{x}_i^+, S^-)$ measures the similarity between the data point $\mathbf{x}_i^+$ and the data set S^-. Similarly, the membership value u_i^- of a data point in S^- can be obtained.

C. *Parameter Selection*

In this approach, the FSVM with the RBF kernel is adopted to classify data with outliers. In the FSVM, the regulation parameter C controls the trade-off between the margin maximization and the amount of misclassification, and the kernel parameter of the RBF, σ^2, controls the capability of the classifier. Since the outliers are associated with low membership values, the parameter C can be set to a sufficient large value such that the FSVM can obtain a smaller misclassification rate for the main body sets. The selection of parameter σ^2 for the RBF kernel is also non-trivial. Recently, several methods were proposed to investigate the selection of Gaussian kernel parameter [21], [22]. In addition, Ying et al., [23] proposed an optimization approach to train SVMs with hybrid kernels such that a superior generalization performance over test data could be obtained, where the parameters of the hybrid kernels are determined by minimizing the upper bound of the VC dimension. According to [23], since only the RBF kernel is adopted in this paper, the objective function becomes

$$\min_{\sigma^2} \quad q(\sigma^2) = R\|\mathbf{w}\| \tag{35}$$

$$\textit{subject to} \quad b_0 = \frac{1}{N_{SV}}\left[\sum_{j\in SV} y_j - \sum_{i,j=SV} y_i\alpha_i K(\mathbf{x}_i, \mathbf{x}_j)\right] \tag{36}$$

and Eq.(8), in which R is the radius of the smallest sphere containing all of the transformed data points. This optimization problem is to minimize the upper bound of the VC dimension through the parameter adjustment of the kernel. More detailed descriptions can be found in [23]. In this paper, the optimization criterion is used to choose the kernel parameter σ^2 for the main body sets of the training data. On the other hand, it should be noted that the ODA technique can approximately represent the actual density shape of the data sets. Thus, the kernel parameter σ^2 can also be directly estimated from the results of outlier detection of S^+ and S^-, and it is defined as

$$\sigma^2 = \frac{1}{2}\left(\tau_b^+ + \tau_b^-\right) \tag{37}$$

where $\tau_b^+ = \beta^+ / \lambda_e^+$ and $\tau_b^- = \beta^- / \lambda_e^-$. In the following section, these two methods for selecting the kernel parameter σ^2 will be adopted and compared.

4. Experimental Results

To illustrate the FSVM of this paper, artificial data and benchmark data are conducted to evaluate the performance. Data sets with different distribution shapes and outliers are first experimented to evaluate the proposed method. The data sets are shown in Figure 4, where Plots (c) and (d) are of the same shape but with different distances between classes. Training samples in data sets S^+ and S^- are indicated by "◊" and "∗" symbols, respectively. Table 1 lists the numbers of samples and outliers in each case. In the experiments, the FSVM with the RBF kernel is adopted to classify the data. The parameters η_1, η_2, η_3, η_4, and C are set

to be 10, 1, 10, 1, and 500, respectively, θ_T^{+-} is set to be θ_T^- for S^-, and the kernel parameter are estimated by Eq. (37) and by minimizing the upper bound of the VC dimension (i.e., Eq.(35)). Figures 5 (a), (d), (g), and (j) show the classification results of the SVM with kernel parameters estimated by Eq. (37), where the black solid curves represent the separating boundaries. The support vectors for each class are marked with red circle, and the color dot curves indicate the equal output levels of the SVM classifier between -1 and +1 with an interval of 0.2. From the classification results of the SVM classifiers, it is clear that the decision surfaces deviate severely from the optimal ones due to the unawareness of outliers.

When the proposed method is adopted to classify the test examples, the corresponding results with the kernel parameter estimated by Eq. (37) are shown in Figures 5 (b), (e), (h), and (k). Table 2 gives the estimated parameter λ_e from the ODA, and the number of the detected outliers. Although the over-fitting problem due to the outliers occurs in Figure 4 (a), the sigmoid fuzzy function can determine the relative importance of data points in the corresponding class, and the obtained decision surfaces are less sensitive to the outliers. In addition, the proposed FSVM method with the kernel parameter determined by minimizing the upper bound of the VC dimension is used to classify the same data sets. The results are shown in Figures 5 (c), (f), (i), and (l). Table 3 lists the values of kernel parameters for the examples estimated by these two methods. It is obvious that the proposed FSVM method can largely reduce the effect of outliers using the kernel parameter σ^2 determined by Eq. (37) or Eq. (35). However, the computational complexity using Eq. (35) is much higher.

To further evaluate the classification and generalization performance of the FSVM, we apply the approach to the banana data, twonorm data and thyroid data from UCI listed in Table 4, where each data set is split into 100 sample sets of training and test set. The parameters η_1, η_2, η_3, and η_4 are set to be 1, 0.5, 1, and 1.2, respectively. θ_T^{+-} is obtained from the ODA for S^-. Then the performance between the SVM and the FSVM is measured by their average error over one hundred partitions of the dataset into training and test sets. For our comparison, the kernel parameter for each training and test set is estimated by Eq. (37) directly, and the robustness test is conducted with C =1, 10, 50, 100, 500, 1000, and δ%=10%, 15%, 20%. Table 5 lists the average test error rates for the SVM and the proposed FSVM while varying C and δ%. It is obvious that the FSVM has better performance in most cases.

For comparison, the same data sets are used to evaluate the SVM, the FSVM using strategy of kernel-target alignment (KT) [24], and the FSVM using strategy of k-NN (k-NN) [24] in which the parameters are the same as stated in [24]. Table 6 lists the test error rates where the results of the proposed FSVM are the best performance in Table 5. For thyroid data set, the FSVM using strategy of k-NN can not improve the performance of SVMs [24]. Thus, we leave blank in Table 6. The simulation results show that the proposed FSVM is very comparable to the FSVM using strategy of kernel-target alignment (KT) and outperforms its counter part - the conventional SVM.

Table 1. Number of data points and outliers for test examples

Example	S^+		S^-	
	# of data points	# of outliers	# of data points	# of outliers
Figure 4 (a)	50	3	50	3
Figure 4 (b)	105	8	100	8
Figure 4 (c)	150	11	150	11
Figure 4 (d)	150	11	150	11

Table 2. Parameters for the proposed FSVM classifier

Example	S^+		S^-	
	λ_e^+	# of estimated outliers	λ_e^-	# of estimated outliers
Figure 4 (a)	4.2	4	4.8	4
Figure 4 (b)	15.5	8	1.6	8
Figure 4 (c)	1.5	11	1.6	11
Figure 4 (d)	1.5	11	1.6	11

Table 3. Kernel parameter for the proposed FSVM classifier

Kernel parameter	Figure 4 (a)	Figure 4 (b)	Figure 4 (c)	Figure 4 (d)
Eq. (37)	0.56	1.27	1.29	1.29
Eq. (35)	1.08	1.17	2.06	2.25

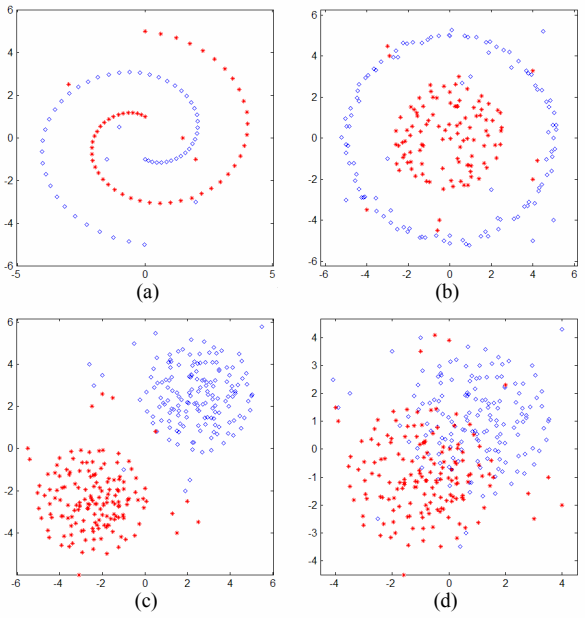


Figure 4. Four test examples.

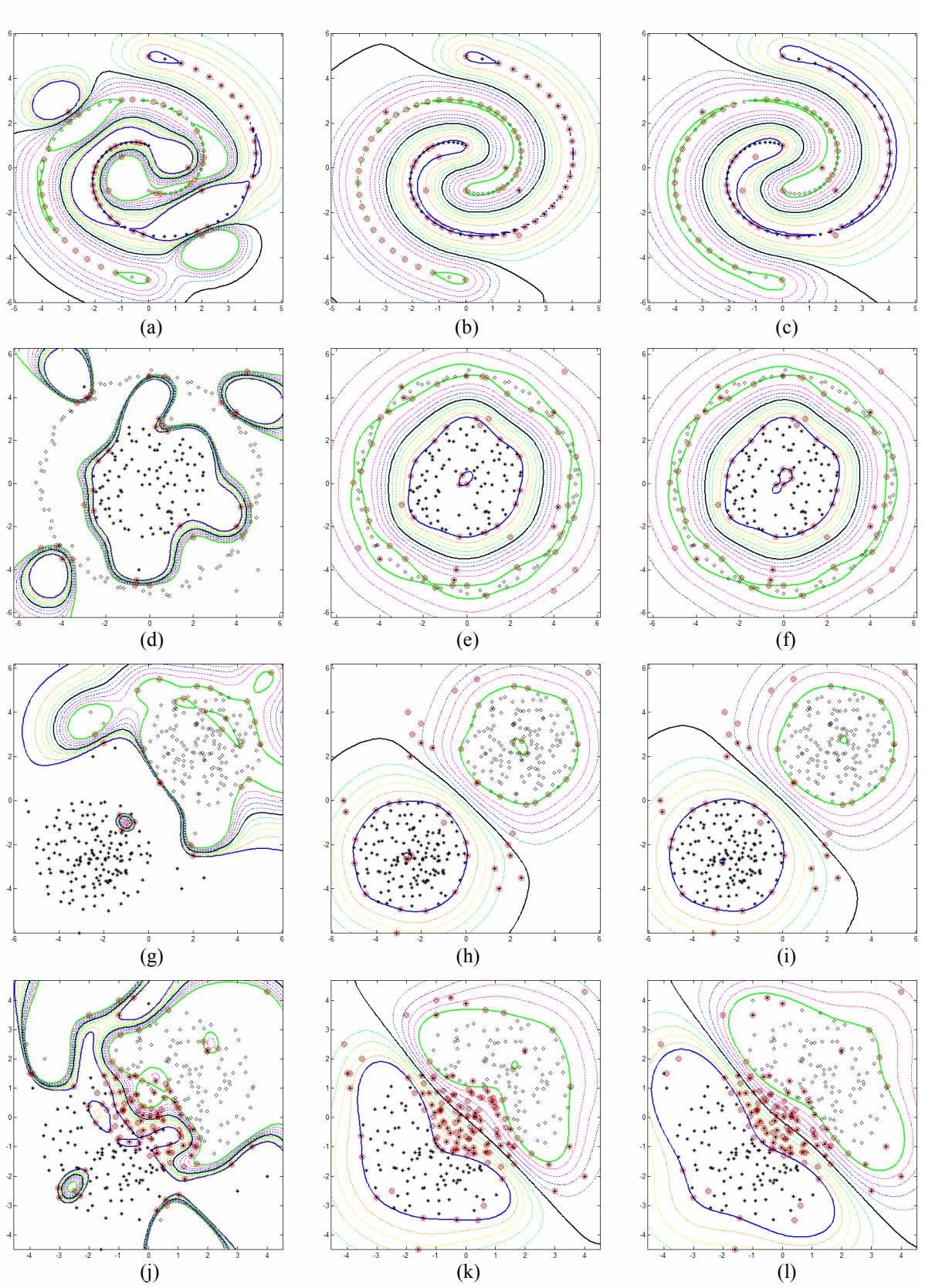


Figure 5. (a), (d), (g) and (j) are the classification results of data points with outliers using the SVM machine. (b), (e), (h) and (k) are the classification results using the proposed FSVM machine with kernel parameter estimated by Eq. (37). (c), (f), (i) and (l) are the classification results using the proposed FSVM machine with kernel parameter determined by minimizing the upper bound of the VC dimension.

Table 4 Feature of benchmark data

Data	# of training patterns	# of test patterns	inputs	classes
Banana	40000	490000	2	2
Twonorm	40000	700000	20	2
Thyroid	14000	7500	5	2

Table 5 Comparison of the average test error rate for the SVM and the proposed FSVM

Data		Banana (%)		Twonorm (%)		Thyroid (%)	
δ%	C	SVM	FSVM	SVM	FSVM	SVM	FSVM
10	1	**10.51**	10.66	2.73	**2.71**	7.21	**6.97**
	10	11.11	**11.10**	3.24	**3.13**	4.91	**4.65**
	50	11.83	**11.77**	3.27	**3.15**	4.29	**4.27**
	100	12.1	**12.06**	3.27	**3.16**	4.16	**4.15**
	500	12.92	**12.89**	3.27	**3.19**	**4.87**	4.88
	1000	13.29	**13.24**	3.27	**3.19**	4.92	**4.92**
15	1	**10.61**	10.75	2.72	**2.70**	6.41	**5.88**
	10	11.32	**11.27**	3.22	**3.10**	4.79	**4.57**
	50	12.03	**11.95**	3.25	**3.13**	4.39	**4.35**
	100	12.35	**12.30**	3.25	**3.14**	4.40	**4.25**
	500	13.27	**13.15**	3.25	**3.17**	**4.84**	4.86
	1000	13.66	**13.57**	3.25	**3.17**	4.83	4.83
20	1	**10.70**	10.86	2.73	**2.70**	5.79	**5.12**
	10	10.52	**10.44**	3.12	**3.08**	4.77	**4.57**
	50	12.24	**12.14**	3.22	**3.11**	4.27	**4.16**
	100	12.61	**12.50**	3.22	**3.11**	4.72	**4.53**
	500	13.62	**13.47**	3.22	**3.14**	4.88	4.88
	1000	14.07	**13.94**	3.22	**3.15**	4.89	**4.89**

Table 6 Comparison of the average test error rate for the SVM, the FSVM using strategy of kernel-target alignment (KT), the FSVM using strategy of k-NN (k-NN), and the proposed FSVM

Data	SVM	KT	k-NN	the proposed FSVM
Banana	11.5	**10.4**	11.4	**10.4**
Twonorm	3.0	**2.4**	2.9	2.7
Thyroid	4.8	4.7	-	**4.2**

5. Conclusion

A systematic method for the two-class classification of data with outliers has been developed in this paper. The essential techniques consist of the outlier detection algorithm (ODA) and the FSVM. In this approach, the main body set for each class is first determined by the ODA. Then, a membership value is assigned to each training sample by the sigmoid fuzzy model. After that, the FSVM with the estimated kernel parameter is adopted to classify the data. Based on the experimental results, the proposed method is shown to be robust against outliers.

6. References

[1] C. Cortes and V. N. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] V. N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[3] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, 1998.

[4] B. Schölkopf, C. Burges, and A. Smola, *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MA: MIT Press, 1999.

[5] Q. Song, W. J. Hu, and W. F. Xie, "Robust support vector machine with bullet hole image classification," *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, pp. 440–448, Nov. 2002.

[6] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using a support vector machine," in *Proc. NNSP*, 1997, pp. 24–26.

[7] H. Drucker *et al.*, "Support vector regression machines," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 9, 1997.

[8] V. Vapnik, S. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *in Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 9, 1997.

[9] C. Chuang, S. Su, J. Jeng, and C. Hsiao, "Robust support vector regression networks for function approximation with outliers," *IEEE Trans. Neural Netwroks*, vol. 13, pp. 1322–1330, Nov. 2002.

[10] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparse approximation," *Neurocomput.*, vol. 48, no. 2, pp. 85-105, 2002.

[11] R. Herbrich and J. Weston, "Adaptive margin support vector machines for classification," in *Proc. 9th ICANN*, vol. 2, pp. 880–885, Sept. 1999.

[12] W. J. Hu, and Q. Song, "An accelerated decomposition algorithm for robust support vector machines." *IEEE Trans. Circuit and System*, vol. 51, pp. 234–240, May 2004.

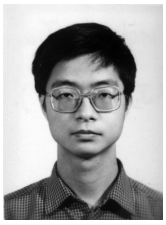
[13] H. P. Huang and Y. H. Liu, "Fuzzy support vector machines for pattern recognition and data mining," *International Journal on Fuzzy Systems,* vol. 4, no. 3, pp. 826-835, Sept. 2002.

[14] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Networks*, vol. 13,

issue. 2, pp. 464-471, Mar. 2002.
[15] Simon Haykin, *Neural Networks: A Comprehensive Foundation,* Second Edition, New Jersey: Prentice-Hall, 1999.
[16] M. Pontil and A. Verri, *Massachusetts Inst. Technol.*, AI Memo no. 1612. Properties of support vector machines, 1997.
[17] V. Barnet and T. Lewis, *Outliers in Statistical Data*, NY: Wiley, 1994.
[18] X. Liu, G. Cheng, and John X. Wu, "Analyzing outlier cautiously," *IEEE Transaction on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 432-437, 2002.
[19] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Systems, Man and Cybernetics*, vol. 24, pp. 1279-1284, 1994.
[20] M. S. Tang, and K. L. Wu, "A similarity-based robust clustering method," *IEEE Trans .Patterrn Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 434–448, April. 2004.
[21] M. Cristianini, J. Shawe-Taylor, and C. Campbell, "Dynamically adapting kernels in support vector machines," *NIPS-98 or NeuroCOLT2 Technical Report Series NC2-TR-1998-017*, Dept. of Engineering Mathematics, Univ. of Bristol, U.K., 1998.
[22] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge Univ. Press, 2000. http://www.suppor-vector.net.
[23] Y. Tan, and J. Wang, "A support vector machine with a hybrid kernel and minimal Vapnik-Chervonenkis dimension," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 4, pp. 385–395, April. 2004.
[24] C. F. Lin and S. D. Wang, "Training algorithm fuzzy for fuzzy support vector machines with noisy data,,"*IEEE XIII Workshop on Neural Networks for Signal Processing*, pp. 517-526, 2003.

Gwo-Her Lee received the B.S. and M.S. degrees in Electrical Engineering from Chung-Cheng Institute of Technology, Taiwan, R.O.C., in 1989 and 1992, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering at National Chung-Hsing University, Taiwan.

Since 1992, he has been an engineer with Aeronautical Research Laboratory, Chung-Shan Institute of Science and Technology. His research interests include flight simulation, 3D graphics, neural networks, fuzzy logic systems, and machine learning.



Jinshiuh Taur received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1987 and 1989, respectively, and the Ph.D. degree in Electrical Engineering from Princeton University, in 1993.

He was a Member of Technical Staff in Siemens Corporate Research, Inc. He is currently a Professor at the National Chung Hsing University, Taiwan, R.O.C. His research interests include neural networks, pattern recognition, computer vision, and fuzzy logic systems.

Dr. Taur received 1996 IEEE Signal Processing Society's Best Paper Award.

C. W. Tao received the B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1984, and the M.S. and Ph.D. degrees in electrical engineering from New Mexico State University, Las Cruces, in 1989 and 1992, respectively. He is currently a Professor with the Department of Electrical Engineering, National I-Lan University, I-Lan, Taiwan. His research interests are on the fuzzy neural systems including fuzzy control systems and fuzzy neural image processing.

Dr. Tao is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. He is listed in *Who's Who in the World*.