International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **7**, July **2013**

ICIC International ©2013 ISSN 1349-4198
pp. **3053–3060**

PROBABILITY FUZZY SUPPORT VECTOR MACHINES

Deqin Yan[1], Xin Liu[2] and Li Zou[1,*]

[1]School of Computer and Information Technology
Liaoning Normal University
No. 1, Liushu South Street, Ganjingzi District, Dalian 116081, P. R. China
*Corresponding author: zoulicn@163.com

[2]School of Mathematics
Liaoning Normal University
No. 850, Huanghe Road, Shahekou District, Dalian 116029, P. R. China

Received April 2012; revised August 2012

Abstract. *In this paper, a model of probability fuzzy support vector machines (PFSVMs) based on the consideration both for fuzzy clustering and probability distributions is proposed. In many applications of traditional support vector machines (SVMs), there are over-fitting problems due to the fact that SVM is sensitive to outliers or noises. In order to solve the problem, the fuzzy support vector machines (FSVMs) model is established. However, in the case that two points are with the same membership, the more information of their influence cannot be carried out by FSVM. The proposed model is based on the consideration that there is not only existing classification distribution but also probability distribution among samples. Experiments show that compared with SVM and FSVM, PFSVM has a better prediction and the classification performance.*
Keywords: Fuzzy support vector machines, Fuzzy clustering, Probability distribution

1. **Introduction.** Support vector machines (SVMs) proposed by Vapnik in the nineties of the 20th century have gained wide acceptance due to their high generalization ability for a wide range of applications and better performance than other traditional learning machines [3], and have drawn much attention on this topic in recent years [4,5,14,15]. For the classification case, SVM has been used for isolated handwritten digit recognition [2,5], speaker identification [1,2], and face recognition [1,18], knowledge-based classification [8], and text categorization [6,12]. Recently, fuzzy theory and technology lead a promising way for the application of SVM in data reconciliation, sound classification, and image de-noising.

However, in the application, standard SVM is sensitive to outliers or noises in the training sample due to over-fitting. To solve this problem, several techniques have been managed. For example, in [4], a central SVM method is proposed to use the class centers in building the SVM. An adaptive margin SVM is developed based on the utilization of adaptive margins for each training pattern [10]. The original input space is mapped to a normalized feature space to increase the stability to noise [9], and a robust support vector machine is proposed aiming at solving the over-fitting problem [16].

Fuzzy support vector machine (FSVM) [13] is another method to solve this problem which is proposed by Lin and Wang. They defined the decision functions according to the membership functions in the directions orthogonal to the hyperplane. In order to decrease the effect of those outliers or noises, FSVM assigns each data point in the training dataset with a membership and sums the deviations weighted by their memberships. If one data point is detected as an outlier, it is assigned with a low membership, so its contribution to

total error term decreases. FSVM can achieve better performance on reducing the effects of outliers than some existing methods.

In many applications, input point may not be appropriately assigned with membership. There is not only existing classification distribution but also probability distribution among samples. Though FSVM can be used to reduce even eliminate the influence of outliers and noises to the whole training model, the probability distribution of sample points is not neglectable. For example, in the case of two points with the same membership, their position and influence are uncertain. For this reason, we propose a model of Probability Fuzzy Support Vector Machines (PFSVMs). The model of PFSVM is based on the idea of building more reasonable classification hyperplane by exploiting more information hidden in data, which is realized by considering both clustering and probability distributions with samples in formulation of PFSVM. Experiments show that compared with SVM and FSVM, Probability Fuzzy Support Vector Machine has a better prediction and the classification performance.

The remainder of this paper is arranged as follows. In Section 2, standard SVM and fuzzy SVM are introduced. The new model of PFSVM is proposed in Section 3. In Section 4, experiment results are carried out to illustrate the advantages of the model. Finally, conclusions are drawn in Section 5.

2. **SVM and Fuzzy SVM.** In this section, we provide a simple introduction about support vector machines and fuzzy support vector machines.

Assuming a training set S is given as $\{x_i, y_i\}$, where $i = 1, \ldots, N$, corresponding class label is $y_i = \{-1, +1\}$. In the linearly separable case, SVM can find a hyperplane to make the largest margin between the two classes without any wrong separated points. This is equivalent to the following quadratic programming (QP) problem:

$$\begin{aligned} &\min \frac{1}{2} w^T w \\ &\text{Subject to: } y_i(w \cdot x_i + b) \geq 1 \end{aligned} \tag{1}$$

To reduce the impact of abnormal (outlier) data points for SVM training model, Lin and Wang proposed a model of fuzzy support vector machines (FSVMs) [13]. In FSVM each sample is given a fuzzy membership in accordance with their importance in class. Fuzzy Support Vector Machine's optimal issue is:

$$\begin{aligned} &\min \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{N} m_i \xi_i \\ &\text{Subject to: } \begin{array}{l} y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0,\ i = 1, 2, \cdots, N \end{array} \end{aligned} \tag{2}$$

where m_i denotes the fuzzy membership of a training sample. The optimal decision function is:

$$\begin{gathered} f(x) = \operatorname{sign}(w \cdot x + b) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i \cdot x) + b \right) \\ 0 \leq \alpha_i \leq m_i C \end{gathered} \tag{3}$$

3. **Model of Probability Fuzzy Support Vector Machines.** Although FSVM can reduce the impact of noise and external samples to the training model, the probability distribution of sample is not taken into account. In FSVM, it is not easy to determine the influence of the points with the same membership value. As illustrate in Figure 1, x_1, x_2 have the same membership to the cluster center v_1, v_2, but have not the same influence.

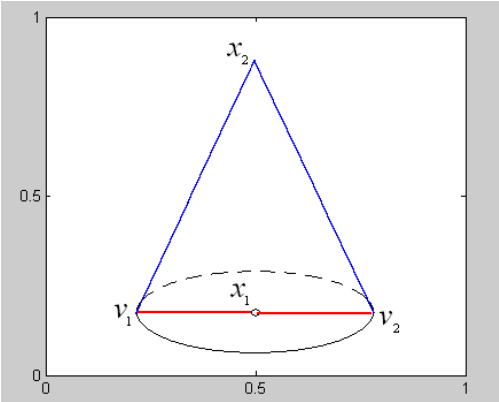

FIGURE 1. The relationship between center and distributed points

In fact, there exists clustering as well as probability distribution features in actual data. Thus, we propose a new model of Probability Fuzzy Support Vector Machine (PFSVM).

3.1. **Probability fuzzy support vector machines (PFSVM).** The objective function PFSVM can be expressed as follows

$$\begin{aligned} &\min \frac{1}{2}w^T w + C\sum_{i=1}^{N}[(u_i^m + t_i^\eta)\xi_i] \\ &\text{Subject to: } \begin{array}{l} y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ i = 1, 2, \cdots, N;\ \xi_i \geq 0 \end{array} \end{aligned} \tag{4}$$

where C, m, η are constant. u_i^m, t_i^η are memberships representing classification and probability distribution respectively.

By applying Lagrange function, the objective Function (4) can be changed into

$$L = \frac{1}{2}w^T w + C\sum_{i=1}^{N}[(u_i^m + t_i^\eta)\xi_i] - \sum_{i=1}^{N}\alpha_i\left[y_i\left(w^T x_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{N}\beta_i\xi_i \tag{5}$$

where α_i, β_i are nonnegative Lagrange multipliers.

We have:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N}\alpha_i x_i y_i = 0 \tag{6}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{7}$$

$$\frac{\partial L}{\partial \xi_i} = C(u_i + t_i) - \alpha_i - \beta_i = 0 \tag{8}$$

Substituting Equation (6), (7) and (8) into Equation (5), we get the QP problem as follows

$$\begin{aligned} &\max Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) + \sum_{j=1}^{N}\alpha_j \\ &\text{Subject to: } \sum_{i=1}^{N}\alpha_i y_i = 0 \quad i = 1, 2, \cdots, N \\ &0 \leq \alpha_i \leq C(u_i^m + t_i^\eta), \quad i = 1, 2, \cdots, N \end{aligned} \tag{9}$$

Accordingly, the KKT conditions are

$$\alpha_i \left[y_i \left(w^T x_i + b \right) - 1 + \xi_i \right] = 0$$
$$\left[C \left(u_i^m + t_i^\eta \right) - \alpha_i \right] \xi_i = 0 \quad i = 1, 2, \cdots N$$

3.2. **Number of class and the cluster center.** Assuming there are c classes and its clustering center is v_i $(i = 1, \cdots, c)$ in the same training class (positive or negative samples), the fuzzy membership of an element x_j to clustering center v_i is $u_{ij} \in [0, 1]$, and probability membership is $t_{ij} \in [0, 1]$. By applying fuzzy probability clustering algorithm [17] we get:

$$v_i = \frac{\sum_{j=1}^{N} \left(u_{ij}^m + t_{ij}^\eta \right) x_j}{\sum_{j=1}^{N} \left(u_{ij}^m + t_{ij}^\eta \right)}, \quad u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \quad t_{ij} = \left(\sum_{k=1}^{n} \left(\frac{||x_j - v_i||}{||x_k - v_i||} \right)^{\frac{2}{\eta-1}} \right)^{-1}$$

$(i = 1, \cdots, c,\ j = 1, \cdots, N)$. u_{ij}, t_{ij} satisfy the following constraints:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad j = 1, \cdots, N$$
$$\sum_{j=1}^{N} t_{ij} = 1, \quad i = 1, \cdots, c$$

The best optimal clustering number c_{best} can be obtained by using partition entropy (PE). PE is defined as:

$$H(U, c) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} |u_{ij} \ln u_{ij}|$$

$$G(T, c) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} |t_{ij} \ln t_{ij}|$$

$$c_{best} = \arg \min_{c=2}^{N-1} \left\{ \min_{U,T} \left\{ \frac{1}{2} \left[H(U, c) \right] + \left[G(T, c) \right] \right\} \right\} \tag{10}$$

3.3. **Assign fuzzy and probability membership for each data point.** Once main body and external data are determined, the fuzzy and probability membership u_i^m can be assigned. In order to obtain a more precise classification hyperplane, we assign different values to main body and external data. For the main body, we defined:

$$u_i = 1 - \frac{\|x_i - \bar{x}\|}{\max_j \|x_j - \bar{x}\|} + \varepsilon, \ x_j \in M \tag{11}$$

where $\|\cdot\|$ is Euclidean distance, ε is a very small positive number, $\bar{x}$ is the center data in main body data set. Therefore, the membership of the main data is defined in scope of $[\varepsilon, 1 + \varepsilon]$.

In this way we can separate the main body from the external data after setting the membership, and reduce the impact of external point to classification.

Assuming that samples obey a certain kind of probability distribution, we calculate the probability of each training sample, and take this value as probability membership t_i of the sample.

3.4. **Theoretic analysis of PFSVM.** We give out the analysis by following theorems:

Theorem 3.1. *In the objective function of PFSVM, the role of t_i^η is to reduce global risk.*

Proof: Let (X, Y) denote a random vector $(X \in R^D, Y \in R)$, $D_N = \left\{(x_i, y_i)_{i=1}^N\right\}$ denote the set of training samples with inputs $x_i \in R^D$, $y_i \in R$. The global risk $R(f)$ of a function $f : R^D \to R$ with respect to a fixed (but unknown) distribution $P_{XY}(x, y)$ is defined as follows ([14])

$$R(f) = \int \ell(y - f(x))\, dP_{XY}(x, y)$$

where $\ell : R \to R$ denote a loss function (e.g., $\ell(e) = e^2$ or $\ell(e) = |e|$).

In the model of PFSVM, $f(x) = \text{sign}[w^T\varphi(x) + b]$. The risk of a training sample x_i with distribution $t_i = P_X^{x_i}(x)$ is $\int \ell(1 - y_i f(x))\, dP_X^{x_i}(x) = c_i t_i$, c_i is constant. So, we can see that the risk is in direct proportion to t_i, and therefore with direct proportion to t_i^η. When x_i is classified correctly, $c_i = 0$, the risk is minimized. The global risk is combined into objective function by t_i^η. That is, in the objective function of PFSVM, the global risk is embodied by t_i^η. By optimizing the objective function of PFSVM, the risk is reduced.

Theorem 3.2. *In the objective function of PFSVM, $(u_i^m + t_i^\eta)$ has direct influence to correct classification.*

Proof: Let $K_i = (u_i^m + t_i^\eta)$, $a_i = a(x_i, w) = (w \cdot x_i + b)$, $\xi_i = [1 - y_i a_i]_+$, where $[u]_+ = \max\{u, 0\}$. Let D denote data set, H denote a probability model, similar to the Bayesian interpretation for SVM in [7], we have

$$p(w|D, \lambda, \mathrm{H}) \propto p(D|w, \mathrm{H})\, p(w|\lambda, H)$$

where λ is parameter.

Assuming that the patterns are independently identically distributed, then

$$p(w|D, \lambda, \mathrm{H}) \propto p(w|\lambda, \mathrm{H}) \prod_i p(y_i|x_i, w, \mathrm{H}) p(x_i)\,. \tag{12}$$

Consider the following probability model:

- Gaussian distribution for w: $p(w|\lambda, \mathrm{H}) \propto \exp\left(-(\lambda/2)\|w\|^2\right)$.
- The probability density function $p(y_i|x_i, w, \mathrm{H})$ for $y_i = \pm 1$ is given by

$$p(y_i|x_i, w, \mathrm{H}) = \frac{\exp\left(-K_i[1 - y_i a_i]_+\right)}{\exp\left(-K_i[1 - a_i]_+\right) + \exp\left(-K_i[1 + a_i]_+\right)}.$$

Substituting these probabilities into (12), we obtain

$$-\log p(w|D, \lambda, \mathrm{H}) = \frac{\lambda}{2}\|w\|^2 - \sum_i \log\left(\frac{\exp\left(-K_i[1 - y_i a_i]_+\right)}{\exp\left(-K_i[1 - a_i]_+\right) + \exp\left(-K_i[1 + a_i]_+\right)}\right) - \sum_i \log p(x_i) + c.$$

where c is constant.

By taking the approximation that $p(y_i|x_i, w, H) \cong \exp\left(-K_i[1 - y_i a_i]_+\right) = \exp(K_i \xi_i)$, we get

$$-\log p(w|D, \lambda, H) = \frac{\lambda}{2}\|w\|^2 + \sum_i K_i \xi_i - \sum_i \log p(x_i) + c.$$

In the equation, the last two terms on the right do not depend on w. Let $\lambda = 1/C$, then performing of PFSVM can be regarded as approximately solving probability equation. So the role of K_i in PFSVM can be evaluated by the probability equation.

From equation

$$p\left(y_i|x_i, w, \mathrm{H}\right) = \frac{\exp\left(-K_i\left[1 - y_i a_i\right]_+\right)}{\exp\left(-K_i\left[1 - a_i\right]_+\right) + \exp\left(-K_i\left[1 + a_i\right]_+\right)}$$

we can see that K_i has direct influence to correct classification of x_i. Let $K_i = (u_i^m + t_i^\eta)$, we get the conclusion of the theorem.

4. **Experimental Analysis and Comparison.** The data used in our experiments are from UCI machine learning database (Table 1) and artificial data which imitates possible conditions happened in practical use.

A. Experiment 1: UCI machine data

We randomly select 3/4 of the data as training set while the remaining data are used for test. With experiments two kinds of kernel (polynomial, RBF) are used. The experiment results are shown in Tables 2 and 3. The accuracy is average of 10 times. In the tables, PFSVM (1) represents performing PFSVM with assumption of samples obeying Gaussian distribution, PFSVM (2) represents performing PFSVM with samples obeying t-distribution.

From the experiment results (shown in Tables 2 and 3) we can see that performance of PFSVM is quite better than that of FSVM and SVM.

Table 1. Data sets

Data set	*Attribute*	*Class*	*Number*
breast	9	2	683
pima	8	2	768
heart	13	2	296
bupa	6	2	345
iris	4	3	150
auto	7	3	392
wine	13	3	178
vehicle	18	4	846
glass	9	7	214
machine	7	8	209

Table 2. Prediction accuracy with polynomial Kernel

Data	*Polynomial*			
	SVM	*FSVM*	*PFSVM (1)*	*PFSVM (2)*
iris	0.9730	0.9459	0.9730	1.000
auto	0.7857	0.7245	0.8061	0.8061
wine	0.8636	0.9318	0.9545	0.9318
vehicle	0.6303	0.6398	0.6872	0.6872
glass	0.3396	0.3962	0.3962	0.358
machine	0.5296	0.5259	0.5851	0.5777

TABLE 3. Prediction accuracy with RBF Kernel

Data	*RBF*			
	SVM	*FSVM*	*PFSVM (1)*	*PFSVM (2)*
iris	0.9730	0.8919	0.9730	0.9730
auto	0.7755	0.7041	0.8061	0.7857
wine	0.8636	0.8636	0.8864	0.9091
vehicle	0.6967	0.6967	0.7488	0.7583
glass	0.3774	0.3774	0.3962	0.4528
machine	0.4615	0.4423	0.4615	0.4615

B. Experiment 2: Artificial data

In order to test the classification ability of PFSVM, artificial data are provided (see Figure 2). The classification results of classification by SVM, FSVM and PFSVM respectively (RBF kernel is used) are shown by Figures 3-5. The percentage shown in brackets is correct rate.

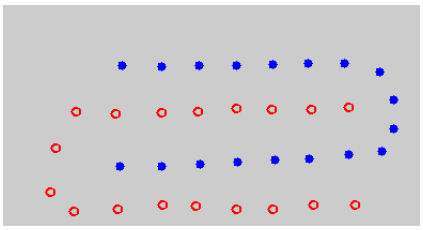

FIGURE 2. Two-class data

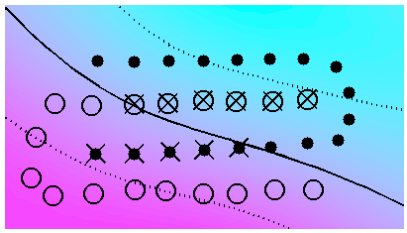

FIGURE 3. Classification of SVM (69.44%)

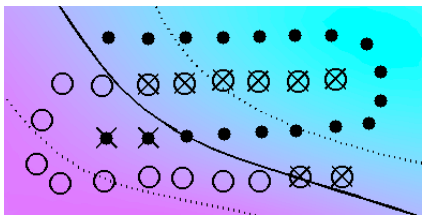

FIGURE 4. Classification of FSVM (72.22%)

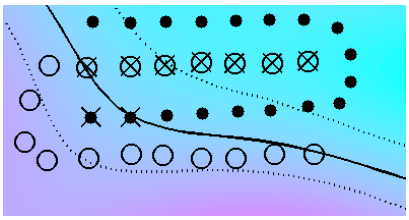

FIGURE 5. Classification of PFSVM (75%)

From classification results of artificial data we can see that the classification accuracy of FSVM increases by nearly 3 percentage points than SVM, PFSVM increases more than two percentage points than FSVM, and it is clear that distance between positive surface and negative surface is narrowed greatly by PFSVM.

5. **Conclusions.** With considering the information of fuzzy and probability distribution in data, a new model of PFSVM is proposed in this paper. The model aims to extend classification ability of SVM and FSVM by exploiting more information hidden in data. In fact, probability distribution of data is non-neglectable in many practical uses. The model of PFSVM provides a tool to make use of the information of training data. The experiments show the algorithm of PFSVM outperforms SVM and FSVM, which improves the generalization ability of SVM and FSVM.

Acknowledgments. This work is partly supported by National Natural Science Foundation of China (Grant Nos. 61105059, 61175055, 61173100), International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61210306079), China Postdoctoral Science Foundation (2012M510815), Liaoning Excellent Talents in University (LJQ2011116), Sichuan Key Technology Research and Development Program under Grant No. 2011FZ00-51, Sichuan Key Laboratory of Intelligent Network Information Processing (SGXZD1002-10) and Key Laboratory of the Radio Signals Intelligent Processing (Xihua University) (XZD0818-09).

REFERENCES

[1] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, Fusion of face and speech data for person identity verification, *IEEE Transactions on Neural Networks*, vol.10, no.5, pp.1065-1074, 1999.

[2] C. J. C. Burges and B. Scholkopf, Improving the accuracy and speed of support vector learning machines, *Advances in Neural Information Processing Systems*, pp.375-381, 1997.

[3] J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, vol.10, no.2, pp.121-167, 1982.

[4] X. G. Zhang, Using class-center vectors to build support vector machines, *Proc. of the 6th IEEE Conf. on Neural Networks and Signal Processing*, pp.3-11, 1999.

[5] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning*, vol.20, pp.273-297, 1995.

[6] K. Crammer and Y. Singer, On the learnability and design of output codes for multiclass problems, *Proc. of the 13th Annual Conf. on Computational Learning Theory*, pp.35-46, 2000.

[7] J. T.-Y. Kwok, The evidence framework applied to support vector machines, *IEEE Transactions on Neural Networks*, vol.11, no.5, pp.1162-1173, 2000.

[8] G. Fung, O. L. Mangasarian and J. Shavlik, Knowledge-based support vector machine classifiers, *Advances in Neural Information Processing Systems*, pp.521-528, 2002.

[9] A. B. A. Graf, A. J. Smola and S. Borer, Classification in a normalized feature space using support vector machines, *IEEE Transactions on Neural Networks*, vol.14, no.3, pp.597-605, 2003.

[10] R. Herbrich and J. Weston, Adaptive margin support vector machines for classification, *Proc. of the 9th Int. Conf. on Artificial Neural Networks*, pp.880-885, 1999.

[11] H. P. Huang and Y. H. Lin, Fuzzy support vector machines for pattern recognition and data mining, *International Journal of Fuzzy System*, vol.4, no.3, pp.826-835, 2002.

[12] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, *Proc. of the 10th European Conf. on Machine Learning*, pp.137-142, 1998.

[13] C. F. Lin and S. D. Wang, Fuzzy support vector machines, *IEEE Transactions on Neural Networks*, vol.13, no.2, pp.464-471, 2002.

[14] V. Vapnik, *Statistical Learning Theory*, Wiley Publishers, New York, 1998.

[15] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch and A. Smola, Input space vs. feature space in kernel-based methods, *IEEE Transactions on Neural Networks*, vol.10, no.5, pp.1000-1017, 1999.

[16] Q. Song, Robust support vector machine with bullet hole image classification, *IEEE Transactions on Systems, Man and Cybernetics*, vol.32, no.4, pp.440-448, 2002.

[17] N. R. Pal, K. Pal and J. C. Bezdek, A mixed C-means clustering model, *Proc. of the IEEE Int. Conf. on Conf. Fuzzy Systems*, pp.11-21, 1997.

[18] E. Osuna, R. Freund and F. Girosi, An improved training algorithm for support vector machines, *Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, pp.276-285, 1997.

[19] D. Chen, Q. He and X. Wang, FRSVM: Fuzzy rough set based support vector machines, *Fuzzy Sets and Systems*, vol.161, no.4, pp.596-607, 2010.