Fuzzy Rough Set Membership based Fuzzy Multiple Kernel Support Vector Machine

Qiang He, Jiaoyang Zhang
School of Science, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China

Abstract: Support vector machines (SVMs) are currently widely used machine learning techniques. SVM is used to construct an optimal hyper-plane that implies an extraordinary generalization capability and good performances. So far, SVMs have already been successfully applied to many real fields. In view of the difficulties in kernel selection and sensitivity to noise, we propose fuzzy rough set membership based fuzzy multiple kernel support vector machine in this paper. The membership degree generalized by fuzzy rough set is introduced to fuzzy multiple kernel support vector machine. It not only avoids the problem of kernel selection, but also improves the robustness to noise. The experimental simulation also validates the feasibility and effectiveness of the method.

Keywords: Multiple Kernel Learning; Support Vector Machine; Fuzzy Rough Set; Fuzzy Membership; Classification

1. Introduction

Support Vector Machine proposed by Vapnik et al. in 1995 is a machine learning algorithm based on statistical learning theory [1,2]. It can improve the generalization ability of the learning machine with the principle of minimizing structural risk [2,3] as well as achieve the minimization of empirical risk and confidence range, which can obtain a good statistical rule in the case of less statistical samples. Support vector machine has been applied in many fields, such as text classification, speech recognition, emotional analysis and regression analysis [4-6] since it is a powerful tool for solving the problems of small sample, nonlinear, high dimension, etc.

However, SVM is sensitive to noise points and outliers. In order to solve this problem, Lin et al. put forward the concept of Fuzzy Support Vector Machine (FSVM) [7, 8]. In other words, which introduced the fuzzy membership into the support vector machine as weight of every sample. FSVM decreases the effect of the noise, improves the classification accuracy. Nowadays FSVM has been applied in many fields including risk prediction, fault diagnosis, handwritten string recognition, etc [9-11].

It is very important for SVM to select kernel function. However, there are no general methods to complete this work. In recent years, multiple kernel learning (MKL) [12-15] which is an important achievement of the kernel method has become a research hot topic in the field of machine learning. Different kernel functions correspond to different similarity expressions. A single kernel function often cannot adequately describe the similarity between the data, especially the similarity between complex data. Hence, the combination of multiple kernels can character the similarity of the data more accurately, and can avoid the problem of kernel function selection.

In this paper, motivated by FSVM and MKL, we propose a fuzzy multiple kernel support vector machine based on fuzzy rough set membership to overcome problem of kernel function selection and sensitivity to noise. Membership evaluated by fuzzy rough set is introduced into multiple kernel support vector machines. Experiments show that the proposed fuzzy multiple kernel SVM (FMKSVM) has better performance than the classical SVM, FSVM and multiple kernel SVM (MKSVM).

The following contents are arranged as follows. Section II briefly reviews some basic notations and definitions in SVM and MKL. Fuzzy multiple kernel SVM based on fuzzy rough set membership is introduced in Section III. The experimental simulation will be given in t Section IV. Finally, Section V concludes the whole paper.

2. Preliminaries

In this section, we will first review the basic contents of SVM, FSVM [1,2,3,7,8,16]. We will then recall the related contents of MKL [12-15,17,18].

2.1. SVM and FSVM

For a dataset $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$, $x\in R^n$ and $y_i\in\{-1,1\}$ for $i=1,2,\ldots,l$. SVM aims to get the optimal separating hyperplane $f(x) = w^{\mathrm{T}}x + b$, which can not only classify the data correctly into two categories, but also maximize the margin between two classes. Where w is the weight vector and $b \in R$ is the threshold value. According to the principle of structural risk minimization, the procedure of searching the optimal hyperplane can be summarized as the following optimization problem

$$\begin{aligned} &\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} x_i \\ &s.t. \ \ y_i\left(w^{\mathrm{T}} \cdot x_i + b\right) \geq 1 - x_i \\ &\quad x_i \geq 0, \ \ i = 1, 2, \mathbf{L}, l \end{aligned} \tag{1}$$

where x_i is the error term, C determines the trade-off between margin maximization and training error minimization.

Using the Lagrange multiplier method transform the above optimization problem into the following duality problem [1,3]:

$$\begin{aligned} &\max \ \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \left\langle x_i, x_j \right\rangle \\ &s.t. \ \sum_{i=1}^{l} a_i y_i = 0, \ \ 0 \leq a_i \leq C, \ \ i, j = 1, 2, \mathbf{L}, l \end{aligned} \tag{2}$$

a_i is the Lagrange multiplier of the sample point x_i, $\langle \cdot, \cdot \rangle$ is the inner product. The decision function $f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} a_i y_i \left\langle x_i, x \right\rangle + b\right)$ can be obtained.

For the nonlinear separable data, by introducing the error term x_i, one can obtain separating hyperplane, but cannot classify all samples correctly.

So SVM was extended to the feature space [1,3]. Through some nonlinear mapping $\Phi : R^n \to H$, all patterns x_i are projected from original space to high dimension feather space. Without any knowledge of the mapping, the optimal separating hyperplane is construct by using the dot product function in the feature space. The dot function is usually called a kernel function. According to Hilbert-Schmidt theorem [2,3], there exists a relationship between the original space and its feature space for the dot product of two points. That is

$$k(x_i, x_j) = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle \tag{3}$$

$k(x_i, x_j)$ is conventionally called a kernel function satisfying the Mercer theorem [2]. Replacing the inner product $\langle x_i, x_j \rangle$ in (1) and (2) with $k(x_i, x_j)$, the optimal separating hyperplane becomes the following form:

$$f(x) = \operatorname{sign}\left(\sum\nolimits_{i=1}^{l} y_i a_i k(x_i, x) + b\right) \tag{4}$$

In uncertain environment, it is difficult for SVM to obtain a satisfactory decision function because data is often disturbed by various factors such as noise. To improve robustness of SVM, in 2002, Liu put forward that a training sample can be assigned a smaller membership degree which can be introduced into the SVM optimization problem model (1-2) when the training sample is identified as a wild point. This kind of SVM is called fuzzy SVM (FSVM) [7,8,16] which corresponds to the following optimization problem:

$$\begin{aligned} &\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} s_i x_i \\ &s.t. \ \ y_i\left(w^{\mathrm{T}} \cdot \Phi(x_i) + b\right) \geq 1 - x_i \\ &\quad x_i \geq 0, i = 1, 2, \mathbf{L}, l \end{aligned} \tag{5}$$

where s_i is the membership generalized by some outlier detecting method, It is clear that s_i plays a weighting role on the x_i in the objective function, so that the noise and outliers have a smaller influence on the resulting hyperplane. This algorithm enhances the robustness of SVM.

The (2) can be converted into its dual form by the Lagrange multiplier method:

$$\begin{aligned} &\max \ \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j k(x_i, x_j) \\ &s.t. \ \sum_{i=1}^{l} a_i y_i = 0, \ \ 0 \leq a_i \leq Cs_i, \ \ i, j = 1, 2, \mathbf{L}, l \end{aligned} \tag{6}$$

The decision function is

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} a_i y_i k(x_i, x) + b\right) \tag{7}$$

Based on the above FSVM algorithm, other scholars also proposed various fuzzy SVM to deal with different specific problem. All of these methods are put forward for some uncertainty in practical problems, they are the improvement and perfection of traditional SVM.

2.2. Multiple kernel learning

There are several kernel functions successfully used in the literature, such as the linear kernel, the polynomial kernel, and the Gaussian kernel. These different kernels may correspond to using different notions of similarity. Selecting the kernel function and its parameters is an important issue in SVM. However, there is not a general method for this work. Generally, a cross-validation procedure is used to choose kernel function, which is high computational complexity. In recent years, multiple kernel learning (MKL) methods have been proposed. MKL combines multiple kernel functions to replace a single kernel function its corresponding parameters.

In Multiple kernel support vector machine (MKSVM), combination of kernel and the calculation of weight is mainly considered. At present, the combination of kernel functions can be divided into the following three types:

Linear combination methods [19] are the most widely used ways currently and have two fundamental categories: unweighted sum (such as the sum of simple kernel functions) and weighted sum, the formula is as follows:

$$k_h\left(x_i, x_j\right)=\sum_{m=1}^{P} h_m k_m\left(x_i^m, x_j^m\right) \tag{8}$$

where h is weight of the kernel, and h is set to 1 in the sum of simple single kernel functions, p is the number of kernel functions.

Nonlinear combination methods [20], such as multiplication, power, and exponentiation, the multiplication of the formula is as follows:

$$k_h\left(x_i, x_j\right)=\prod_{m=1}^{P} k_m\left(x_i^m, x_j^m\right) \tag{9}$$

Data-dependent combination methods [17], this method will be assigned a specific weight for each sample, by which the local distribution of data can be known. Therefore, different data regions have different combination of kernels.

The calculation methods of weight coefficients include five categories: fixed rules [18], heuristic approaches [21], optimization approaches [22], Bayesian approaches [17] and boosting approaches [23].

It is worth noting that the conclusion of section 2.1 is still valid if we substitute $k_h\left(x_i, x_j\right)$ for $k(x_i, x_j)$. And the decision function is changed to

$$f\left(x\right)=\operatorname{sgn}\left(\sum_{i=1}^{l} a_i y_i k_h\left(x_i, x\right)+b\right) \tag{10}$$

3. Fuzzy Rough Set Membership based Fuzzy Multiple Kernel Support Vector Machine

Aiming at the difficulties in kernel selection and the sensitivity to noise, this paper introduces the sample membership degree obtained by fuzzy rough set method to multiple kernel support vector machine, and proposes fuzzy rough set membership based fuzzy multiple kernel support vector machine (FMKSVM).

Rough set (RS) was originally proposed by Pawlak [24] as a mathematical approach to handle uncertainty in data analysis. However, RS can only deal with databases with symbolic values. In order to handle databases with real number values, RS and fuzzy sets are combined together and so fuzzy rough sets (FRS) are derived by Dubois and Prade [25,26]. Since then, FRS have a fast development [27-32]. FRS has already been successfully applied to many real fields [33-35]. In this paper, we use lower approximation operator of Gaussian kernel based FRS as membership [36,37]. For $x \in A$, $s=\inf_{x' \notin A} \sqrt{1-\left(k\left(x, x'\right)\right)^2}$. Where s is the membership that x belongs to A, and $k(x, x')$ is Gaussian Kernel.

Unweighted linear combination is adopted in this paper. The optimization problem of FMKSVM is changed from (5) to

$$\begin{aligned} &\min \ \frac{1}{2}\left\|w_h\right\|^2+C\sum_{i=1}^{l} s_i x_i \\ &s.t. \ y_i\left(w_h{}^{\mathrm{T}} \cdot \Phi_h\left(x_i\right)+b\right) \geq 1-x_i \\ &x_i \geq 0, i=1,2,\mathbf{L},l \end{aligned} \tag{11}$$

With the introduction of Lagrange multiplier, the dual problem can be transformed from (5) to:

$$\begin{aligned} &\max \ \sum_{i=1}^{l} a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j k_h\left(x_i, x_j\right) \\ &s.t. \ \sum_{i=1}^{l} a_i y_i=0, \ 0 \leq a_i \leq C s_i, \ i, j=1,2,\mathbf{L},l \end{aligned} \tag{12}$$

The decision function $f\left(x\right)=\operatorname{sgn}\left(\sum_{i=1}^{l} a_i y_i k_h\left(x_i, x\right)+b\right)$ is obtained.

4. Experimental Simulation

The UCI database, a database used by the University of California at Irvine for machine learning, is a commonly used standard test data set. In order to verify the feasibility and effectiveness of the proposed method, 7 data sets are selected in the UCI database, as shown in Table 1. Since this article only considers the two-class classification problem, the class 2 and 3 are considered as a class for the wine data set.

Table 1. Data Information

Number	Dataset name	Number of samples	Number of attributes	Number of categories
1	breastcancer	683	10	2
2	bupa	345	7	2
3	ionosphere	351	34	2
4	pima	768	9	2
5	sonar	208	61	2
6	wdbc	569	31	2
7	wine	178	14	3

Experiments run on a PC (CPU: 2.60GHz memory 4.00GB), the operating system is Windows 8.1, and the experimental tool is Matlab R2014b. In the experiment, all data are normalized. The kernel function uses polynomial kernel and Gaussian kernel. From the comparison of experimental results, the results are better in most cases when the two kernel parameters are 2, as shown in Figure 1. So the parameters of the two kernel functions are set to 2. Given the penalty factor $C=100$. The experiment compares the classification performance of fuzzy rough set membership based multiple kernel fuzzy support vector machine (MFSVM) and classical support vector machine (SVM), fuzzy support vector machine (FSVM) and multiple kernel support vector machine (MSVM). The 10 fold cross validation method were used in the experiment.

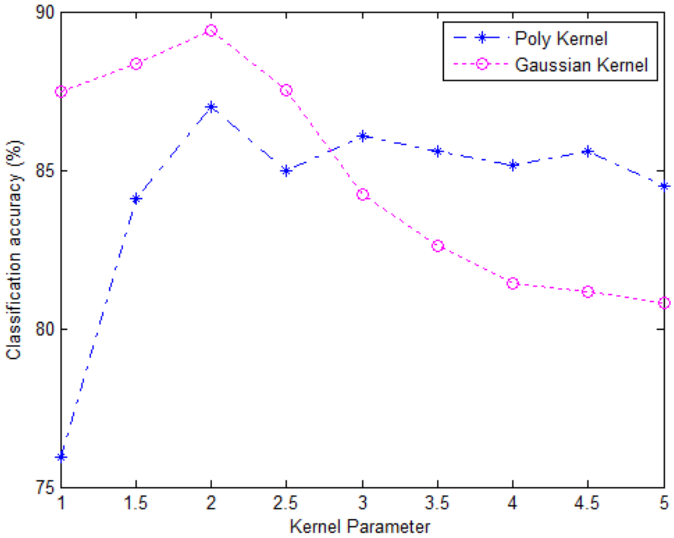

Figure 1. Classification Accuracies of Sonar Versus the Values of the Kernel Parameter

The results of the experiment are shown in Table 2. It can be seen that the FMKSVM method proposed in this article has higher classification accuracy on most databases. This also validates the effectiveness of the method.

To further verify the performance of proposed FMKSVM in noisy environments, then we randomly select some of the cases in all seven databases to change their class labels. Table 3 and table 4 are the results of the test accuracy of the 4 algorithms when the noise ratio is 10% and 30% respectively.

From the experimental results, it can be seen that the classification accuracy of FMKSVM method in 6 databases is higher than that of the other three methods when 10% noise is added. On sonar, FMKSVM and FSVM have the same and the highest classification accuracy. The FMKSVM has the highest classification accuracy on all databases although the classification accuracy of the Table 4 methods is relatively lower than that of Table 3 at a noise ratio of 30%. This result further verifies that the proposed FMKSVM not only avoids the problem of kernel selection, but also has strong anti-noise capability. Thus, FMKSVM is not only feasible, but also has a wider range of applications.

Table 2. Comparison of Classification Accuracy without Noise

Number	Dataset	Test accuracy(%)			
		SVM	FSVM	MKVM	FMKSVM
1	breastcancer	94.60	95.47	94.75	95.47
2	bupa	72.62	71.29	72.90	71.86
3	ionosphere	87.48	88.33	87.77	89.18
4	pima	76.03	76.29	77.06	76.82
5	sonar	86.99	86.99	87.27	87.27
6	wdbc	96.14	97.54	96.31	97.72
7	wine	96.67	97.78	96.67	97.78

Table 3. Comparison of Classification Accuracy with 10% Noise

Number	Dataset	Test accuracy(%)			
		SVM	FSVM	MKVM	FMKSVM
1	breastcancer	87.40	87.55	86.55	87.87
2	bupa	65.52	56.62	65.67	66.20
3	ionosphere	75.48	79.75	74.91	80.61
4	pima	71.49	71.87	71.73	72.90
5	sonar	74.94	74.94	75.41	75.41
6	wdbc	82.95	84.54	84.36	86.82

7	wine	87.64	88.19	84.24	88.61

Table 4. Comparison of Classification Accuracy with 30% Noise

Number	Dataset	Test accuracy(%)			
		SVM	**FSVM**	**MKVM**	**FMKSVM**
1	breastcancer	68.53	68.38	68.82	69.55
2	bupa	54.62	48.90	54.05	56.14
3	ionosphere	60.67	59.53	60.67	62.68
4	pima	58.97	59.75	59.64	61.57
5	sonar	58.97	59.87	60.35	61.45
6	wdbc	64.69	65.74	64.33	66.09
7	wine	63.85	66.15	67.69	68.46

4 . Conclusions

This paper introduced fuzzy rough set membership to multiple kernel support vector machine, and FMKSVM was established. The problem of kernel selection is avoided, and enhanced robust to noise through FMKSVM. The experimental results show that this method combines the advantages of fuzzy support vector machines and multiple kernel learning, and improves the classification performance of FSVM. It could be interesting to investigate how to improve the calculation of sample membership and weight of kernels for FMKSVM.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 61473111, and Research Foundation of Beijing University of Civil Engineering and Architecture (KYJJ2017017).

References

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1999: 69-110.

[2] Vapnik V N. Statistical Learning Theory[M]. New York: A Wiley-Interscience Publication, 1998

[3] Deng N Y, Tian Y J. A New Method in Data Mining - Support Vector Machine[M]. Beijing: Science Press, 2004.

[4] Mahdi G V, Mojtaba T, Puneet M, Zahra I. Temperature Modulation of Electronic Nose Combined with Multi-Class Support Vector Machine Classification for Identifying Export Caraway Cultivars[J]. Postharvest Biology and Technology, 2018, (138): 134-139.

[5] Chen W, Zhang J F, Gao M Y, Shen G X. Performance Improvement of 64-QAM Coherent Optical Communication System by Optimizing Symbol Decision Boundary Based on Support Vector Machine[J]. Optics Communications, 2018, 410: 1-7.

[6] Bissan G Joe N S. High Dimensional Data Classification and Feature Selection Using Support Vector Machines[J]. European Journal of Operational Research, 2018, 265(3): 993-1004.

[7] Lin C F, Wang S D. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

[8] Huang H P, Liu Y H. Fuzzy Support Vector Machines for Pattern Recognition and Data Mining[J]. Intl.j.of Fuzzy Systems, 2002, 4(3):826-835.

[9] Fan Q, Wang Z, Li D D, Gao D Q, Zha H Y. Entropy-Based Fuzzy Support Vector Machine for Imbalanced Datasets[J]. Knowledge-Based Systems, 2017, (115): 87-99.

[10] Hang J, Zhang J Z, Cheng M. Application of Multi-Class Fuzzy Support Vector Machine Classifier for Fault Diagnosis of Wind Turbine[J]. Fuzzy Sets and Systems, 2016, 297: 128-140.

[11] Liu Y, Zhang B L, Chen B, Yang Y D. Robust Solutions to Fuzzy One-Class Support Vector Machine[J]. Pattern Recognition Letters, 2016, 71: 73-77.

[12] Bach F R, Lanckriet G R G, Jordan M I. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm[C]// Proc. International Conference on Machine Learning. 2004:6.

[13] Zhao Y G, Song Z, Z Feng, Shao L Learning a Multiple Kernel Similarity Metric for kinship verification[J]. Information Sciences, 2018, 430–431:247-260.

[14] Sajjad K S, Poorya Z M, Hamid B. Detection of Evolving Concepts in Non-Stationary Data Streams: A Multiple Kernel Learning Approach[J]. Expert Systems with Applications, 2018, 91: 187-197.

[15] Mehmet G, Ethem A. Multiple Kernel Learning Algorithms[J]. Journal of Machine Learning Research 2011, 12: 2211-2268.

[16] Lin C F, Wang, S D. Training Algorithms for Fuzzy Support Vector Machines with Noisy Data[J]. Pattern Recognition Letters, 2004, 25(14):1647-1656.

[17] Christoudias C M, Urtasun R, Darrell T. Bayesian Localized Multiple Kernel Learning[J]. Univ California Berkeley, 2009, 7(2006):1531-1565.

[18] Ben-Hur A, Noble W S. Kernel Methods for Predicting Protein-Protein Interactions[J]. Bioinformatics, 2005, Suppl 1:i38-46.

[19] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods[M]. Printed in the United Kingdom at the University Press, 2000.

[20] Diego I M D, Muñoz A, Moguerza J M. Methods for the Combination of Kernel Matrices within A Support Vector Framework[J]. Machine Learning, 2010, 78(1-2):137-174.

[21] Qiu S, Lane T. A Framework for Multiple Kernel Support Vector Regression and Its Applications to siRNA Efficacy Prediction[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2009, 6(2):190.

[22] Varma M, Ray D. Learning the Discriminative Power-Invariance Trade-Off[C]// IEEE, International Conference on Computer Vision. IEEE, 2007:1-8.

[23] Du H Y. Research of Reduced Multiple Kernel Support Vector Machine[D]. Beijing Jiaotong University, 2015.

[24] Pawlak Z. Rough Sets[M]. Kluwer Academic Publishers, 1991.

[25] Dubois D, Prade H. Rough Fuzzy Sets and Fuzzy Rough Sets[J]. International Journal of General Systems[J], 1990, 17(2-3):191-209.

[26] Dubois D, Prade H. Putting Rough Sets and Fuzzy Sets Together[M]// Intelligent Decision Support. Springer Netherlands, 1992:203-232.

[27] Bhatt R B, Gopal M. On Fuzzy-Rough Sets Approach to Feature Selection[J]. Pattern Recognition Letters, 2005, 26(7):965-975.

[28] Chen D G, Tsang E C C, Zhao S Y. An Approach of Attributes Reduction Based on Fuzzy T L, Rough Sets[C]// IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2007:486-491.

[29] Chen D, Wang X, Zhao S. Attribute Reduction Based on Fuzzy Rough Sets[M]// Rough Sets and Intelligent Systems Paradigms. Springer Berlin Heidelberg, 2007:381-390.

[30] Chen D, Tsang E C C, Yeung D S, et al. The Parameterization Reduction of Soft Sets and Its Applications[J]. Computers & Mathematics with Applications, 2005, 49(5-6):757-763.

[31] Hu Q, Yu D, Xie Z. Information-Preserving Hybrid Data Reduction Based on Fuzzy-Rough Techniques[J]. Pattern Recognition Letters, 2006, 27(5):414-423.

[32] Tsang E C C, Chen D, Yeung D S, et al. Attributes Reduction Using Fuzzy Rough Sets[J]. IEEE Transactions on Fuzzy Systems, 2008, 16(5):1130-1141.

[33] Hassanien A E. Fuzzy Rough Sets Hybrid Scheme for Breast Cancer Detection[J]. Image & Vision Computing, 2007, 25(2):172-183.

[34]]Liu M, Chen D, Wu C, et al. Fuzzy Reasoning Based on A New Fuzzy Rough Set And Its Application to Scheduling Problems[J]. Computers & Mathematics with Applications, 2006, 51(9):1507-1518.

[35] Sarkar M. Ruggedness Measures of Medical Time Series Using Fuzzy-Rough Sets And Fractals[J]. Pattern Recognition Letters, 2006, 27(5):447-454.

[36] Chen D G, Yang Y Q, Wang H. Granular Computing Based on Fuzzy Similarity Relations[J]. Soft Computing, 2011, 15(6):1161-1172.

[37] He Q, Wu C. Membership Evaluation and Feature Selection for Fuzzy Support Vector Machine Based on Fuzzy Rough Sets[J]. Soft Computing, 2011, 15(6):1105-1114.