# 一种基于粗糙间隔的模糊支持向量机

李 凯[1],卢霄霞[2]
(1.河北大学数学与计算机学院,河北保定 071002;2.包头供电局电能计量中心,内蒙古包头 014030)

**摘 要:** 以模糊支持向量机(FSVM)为基础,同时考虑样本在间隔中的位置对决策超平面的影响,提出了基于粗糙间隔的模糊支持向量机(RFSVM).通过计算各个数据点的模糊隶属度,并利用最大化粗糙间隔方法,对具有隶属度的数据进行训练以获得决策超平面.在此算法中,位于下间隔中的训练点比边界域中的训练点具有较大的惩罚值,以便更好地减少噪声或野点对超平面的影响.利用选择的标准数据集对几种不同算法进行了实验比较,结果表明了RFSVM算法的有效性.

**关键词:** 模糊支持向量机; 粗糙间隔; 分类; 正确率
**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0372-2112 (2013) 06-1183-05
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.06.021

## A Rough Margin Based Fuzzy Support Vector Machine

LI Kai[1], LU Xiao-xia[2]
(1. *School of Mathematics and Computer, Hebei University, Baoding, Hebei* 071002, *China*;
2. *Power Measurement Center of Baotou Power Supply Bureau, Baotou, Inner Mongolia* 014030, *China*)

**Abstract:** Based on fuzzy support vector machine(FSVM), we presented a rough margin based fuzzy support vector machine (RFSVM) by introducing the effects of positions of training samples in the margin on decision hyper-plane in this paper. After computing the degree of fuzzy membership of each training point, we used these data for training to obtain the decision hyper-plane by maximizing rough margin′s method. In this algorithm, points in the lower margin have major penalty than those in the boundary. We compared RFSVM with other support vector machine algorithms on several benchmark datasets. Experimental results show that RFSVM is effective and feasible.

**Key words:** fuzzy support vector machine(FSVM); rough margin; classification; accuracy

## 1 引言

支持向量机(Support Vector Machine, SVM)最初是由Vapnik[1,2]等提出的一种用于解决二分类问题的学习算法,它在小样本数据集中显示出特有的优势,并且通过引入核函数,将原始空间中的非线性问题转化为特征空间的线性问题来求解.最近几年,支持向量机引起越来越多学者和专家的关注与研究[3].支持向量机的理论基础是VC维和结构风险最小化原理,其目标是寻找一个最优超平面将两类样本在最大化间隔情况下分开.

可以看到,在传统的支持向量机中,所有的数据点对超平面的作用是相同的,而当数据点不能完全被确定为某一类时,此时所产生的分类面往往不是最优超平面.为此,研究人员提出了模糊支持向量机(Fuzzy support Vector Machine, FSVM)[4,5],在该方法中,通过为每个训练点赋予一个模糊隶属度,以减少噪声或野点对最优超平面的影响.之后,Wang[6]等将FSVM应用到信用评估领域,并针对信用问题提出了新的模糊支持向量机模型.

为了解决支持向量机中的过学习问题, Zhang等[7]将粗糙集引入到ν-SVM[8]中,提出了基于粗糙间隔的支持向量机(Rough Margin based Support Vector Machine, RMSVM),但该方法只考虑了样本点对最优分类超平面具有相同作用的情形.后来,许多学者不断将粗糙理论应用于支持向量机中,以解决分类或回归问题,例如,Xu[9]将粗糙理论应用于支持向量回归问题,提出了基于粗糙间隔的线性ν-SVR; Hsiao等[10]将粗糙方法应用于支持向量网络中,以解决函数逼近问题; Chen等[11]提出了基于模糊粗糙集的支持向量机.通过分析研究模糊支持向量机与粗糙间隔支持向量机方法,本文提出了一

收稿日期:2010-11-15;修回日期:2012-12-14
基金项目:国家自然科学基金(No.61073121);河北省自然科学基金(No.F2012201014)

种基于粗糙间隔的模糊支持向量机 RFSVM(Rough margin based Fuzzy Support Vector Machine),通过引入上下间隔且给每个样本赋予不同的模糊隶属度,以此来提高支持向量机的泛化性能.

## 2 基于粗糙间隔的模糊支持向量机(RFSVM)

在粗糙间隔的模糊支持向量机中,同时考虑了训练样本的模糊隶属度和训练样本在粗糙间隔中的位置对最优超平面的影响.需提及的是,在 RFSVM 中,使用了两个间隔且给予不同的惩罚系数,即上间隔 $\rho_u$ 和下间隔 $\rho_l(\rho_u>\rho_l)$,位于下间隔内的训练样本所对应的区域对应于粗糙集中的正域,且此区域的数据为噪声或野点;位于上间隔之外的训练点所对应的区域为负域,此区域的数据点是被正确划分的;而位于上下间隔之间的数据点所对应的区域是边界域,此区域的数据点可能为噪声或野点.

假设训练集为$(\boldsymbol{x}_1,y_1,s_1),(\boldsymbol{x}_2,y_2,s_2),\cdots,(\boldsymbol{x}_l,\boldsymbol{y}_l,s_l)$,其中$(\boldsymbol{x}_i,y_i)\in\mathbf{R}^N\times\{-1,+1\}$,$i=1,2,\cdots,l$,并且每个样本点的模糊隶属度为 $s_i(0<s_i\leqslant 1)$,参数 $\xi_i$ 用于度量支持向量机中的错分,而 $s_i\xi_i$ 表示对错分加权后的度量,$\delta\geqslant 1$ 用于设定粗糙间隔,则 RFSVM 的原始问题表达如下:

$$\begin{aligned}\min_{w,b,\xi,\xi',\rho_l,\rho_u}\ &\frac{1}{2}\|\boldsymbol{w}\|^2-\nu\rho_l-\nu\rho_u+\frac{1}{l}\sum_{i=1}^{l}s_i\xi_i+\frac{\delta}{l}\sum_{i=1}^{l}s_i\xi_i'\\ \text{s.t.}\quad &y_i(\langle\boldsymbol{w},\varphi(\boldsymbol{x}_i)\rangle+b)\geqslant\rho_u-\xi_i-\xi_i',\\ &0\leqslant\xi_i\leqslant\rho_u-\rho_l,\xi_i'\geqslant 0,\rho_l\geqslant 0,\\ &\rho_u\geqslant 0,i=1,2,\cdots,l.\end{aligned}\tag{1}$$

其中 $\nu\in[0,1]$和 $\delta>1$ 是正则化参数,$\xi'_i$ 和 $\xi_i$ 是松弛变量,$\rho_l$ 和 $\rho_u$ 分别是构造边界域的内外墙.可以看到,当 $\delta=1$ 时,则原始优化问题等价于 $\nu$-FSVM.在 RFSVM 中,同时引入松弛变量 $\xi'_i$ 和 $s_i$,并且给予其惩罚为 $\xi_i$ 的 $\delta$ 倍,其中 $\xi'_i>0$ 表示其对应的数据点在下间隔内,即在学习最优超平面的过程中,位于下间隔中的训练点比位于边界的训练点给予较大的惩罚.为了求解此优化问题,构造 Lagrange 函数如下:

$$\begin{aligned}L=&\frac{1}{2}\|\boldsymbol{w}\|^2-\nu\rho_l-\nu\rho_u+\frac{1}{l}\sum_{i=1}^{l}s_i\xi_i+\frac{\delta}{l}\sum_{i=1}^{l}s_i\xi_i'\\ &-\sum_{i=1}^{l}\alpha_i(y_i(\langle\boldsymbol{w},\varphi(\boldsymbol{x}_i)\rangle+b)-\rho_u+\xi_i+\xi_i')\\ &-\sum_{i=1}^{l}\beta_i\xi_i-\sum_{i=1}^{l}\lambda_i(\rho_u-\rho_l-\xi_i)-\sum_{i=1}^{l}\eta_i\xi_i'\\ &-\theta_1\rho_l-\theta_2\rho_u\end{aligned}\tag{2}$$

其中 Lagrange 乘子满足 $\alpha_i\geqslant 0,\beta_i\geqslant 0,\lambda_i\geqslant 0,\ \eta_i\geqslant 0,\theta_1\geqslant 0,\theta_2\geqslant 0$.通过计算相应的偏导数并令其为 0,则有如下的结果:

$$\begin{aligned}&\frac{\partial L}{\partial\boldsymbol{w}}=\boldsymbol{w}-\sum_{i=1}^{l}\alpha_i y_i\varphi(\boldsymbol{x}_i)=0\\ &\frac{\partial L}{\partial b}=-\sum_{i=1}^{l}\alpha_i y_i=0\\ &\frac{\partial L}{\partial\xi_i}=\frac{s_i}{l}-\alpha_i-\beta_i+\lambda_i=0\\ &\frac{\partial L}{\partial\xi_i'}=\frac{\delta s_i}{l}-\alpha_i-\eta_i=0\\ &\frac{\partial L}{\partial\rho_l}=-\nu+\sum_{i=1}^{l}\lambda_i-\theta_1=0\\ &\frac{\partial L}{\partial\rho_u}=-\nu+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}\lambda_i-\theta_2=0\end{aligned}\tag{3}$$

根据 KKT 条件可得:

$$\begin{aligned}&\alpha_i(y_i(\langle\boldsymbol{w},\varphi(\boldsymbol{x}_i)\rangle+b)-\rho_u+\xi_i+\xi_i')=0\\ &\lambda_i(\rho_u-\rho_l-\xi_i)=0,\beta_i\xi_i=0\\ &\eta_i\xi_i'=0,\theta_1\rho_l=0,\theta_2\rho_u=0\end{aligned}\tag{4}$$

将上述等式带入 Lagrange 函数式(2),可得如下的 Wolf 对偶问题:

$$\begin{aligned}\min_{\alpha}\ &\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\\ \text{s.t.}\ &\sum_{i=1}^{l}\alpha_i y_i=0\\ &0\leqslant\alpha_i\leqslant\frac{\delta s_i}{l}\\ &\sum_{i=1}^{l}\alpha_i\geqslant 2\nu,\end{aligned}\tag{5}$$

在求出对偶问题的最优解$\overline{\alpha_i}$后,根据$\overline{\alpha_i}$的值决定训练点在粗糙间隔中的位置.根据 KKT 条件,可以知道:当$\overline{\alpha_i}=0$时,数据点 $x_i$ 位于粗糙间隔(即上间隔)外且满足 $y_i(\langle\overline{\boldsymbol{w}},\varphi(\boldsymbol{x}_i)\rangle+\overline{b})>\rho_u$,即数据点被正确地分类且位于负域内;当$\overline{\alpha_i}>0$时,此时的数据点称为支持向量,且数据点位于正域或边界域中;当 $0<\overline{\alpha_i}<\frac{s_i}{l}$时,位于超平面的上间隔之上的数据点满足 $y_i(\langle\overline{\boldsymbol{w}},\varphi(\boldsymbol{x}_i)\rangle+\overline{b})=\rho_u$;当$\overline{\alpha_i}=\frac{s_i}{l}$时,位于粗糙间隔(上间隔之内和下间隔之外)之间的数据点满足 $y_i(\langle\overline{\boldsymbol{w}},\varphi(\boldsymbol{x}_i)\rangle+\overline{b})=\rho_u-\xi_i$,其中,$\xi_i>0$.当$\frac{s_i}{l}<\overline{\alpha_i}<\frac{\delta s_i}{l}$时,位于下间隔边界上的数据点满足 $y_i(\langle\overline{\boldsymbol{w}},\varphi(\boldsymbol{x}_i)\rangle+\overline{b})=\rho_l$;当$\overline{\alpha_i}=\frac{\delta s_i}{l}$时,位于下间隔内的数据点是被错分的,且满足 $y_i(\langle\overline{\boldsymbol{w}},\varphi(\boldsymbol{x}_i)\rangle+\overline{b})=\rho_u-\xi_i',\xi_i'>0$.

通过求解 RFSVM 的对偶问题,可以获得最优解$\overline{\alpha}(i=1,2,\cdots,l)$,从而得到如下的决策函数:

$$f(\boldsymbol{x})=\operatorname{sgn}\Big(\sum_{i\in\mathrm{RSV}}\overline{\alpha}y_i K(\boldsymbol{x},\boldsymbol{x}_i)+\overline{b}\Big)\tag{6}$$

其中 RSV 表示$\overline{\alpha}>0$时的数据索引集,而

$$\overline{b}=-\frac{1}{2}\sum_{i\in \mathrm{RSV}}\overline{\alpha}_i y_i(K(\boldsymbol{x}_i,\boldsymbol{x}_j)+K(\boldsymbol{x}_i,\boldsymbol{x}_k)),$$

$$i\in\{j\mid\overline{\alpha}_j\in(0,\frac{s_j}{l}),y_j=+1\},$$

$$k\in\{j\mid\overline{\alpha}_j\in(0,\frac{s_j}{l}),y_j=-1\}$$

或 $$i\in\{j\mid\overline{\alpha}_j\in(\frac{s_j}{l},\frac{\delta s_j}{l}),y_j=+1\},$$

$$k\in\{j\mid\overline{\alpha}_j\in(\frac{s_j}{l},\frac{\delta s_j}{l}),y_j=-1\}.$$

从式(5)的限制条件$0\leqslant\alpha_i\leqslant\frac{\delta s_i}{l}$和不等式$\sum_{i=1}^{l}\alpha_i\geqslant 2\nu(\rho_l>0)$可知,如果所有的训练点位于正域内(即下间隔线之间),则有$\sum_{i=1}^{\mathrm{Enu}}\frac{\delta s_i}{l}=2\nu$(Enu 表示间隔中被错分的训练点数),由此可得$\frac{2\nu l}{\delta s_{\max}}<\mathrm{Enu}<\frac{2\nu l}{\delta s_{\min}}$,其中 $s_{\min}$和 $s_{\max}$分别表示最小模糊隶属度$\min_{i\in s_i}(s_i)$和最大模糊隶属度$\max_{i\in s_i}(s_i)$.注意到,当 $\delta=1$ 时,$\mathrm{Enu}<\frac{2\nu l}{s_{\min}}$,即在上间隔内被错分的数据点的个数,因而 $\nu$、$\delta$ 和 $s_i$ 共同控制粗糙间隔中的间隔错分数和边界宽度.通过以上的分析与推导可知,RFSVM 算法的时间复杂性为 $O(l^3)$.

## 3　实验结果及分析

为了表明 RFSVM 的有效性,选择了 UCI[12],Statlog[13]和 TKH96a[14]等 12 个两类问题的数据集,其中 wine 数据集为 3 类问题,但在实验研究中将其视为二类问题,即将第一与第三类视为一类,第二类视为另一类.实验中采用随机选择方法且对每种方法都重复实验 10 次.首先随机选择数据集中 70% 的样本作为训练集,剩余的 30% 作为测试集.实验中使用的参数 $C$ 分别设定为 10 和 100,$\nu$ 的取值在 0.3 – 0.6 之间,$\delta$ 的取值在 3.0 – 15.0 之间,核函数为 Gaussian 核且其核参数 $\gamma=1$,模糊隶属度的计算采用了基于样本点和其类中心的距离方法.

表 1 给出了不同支持向量机算法的实验结果.可以看到,RFSVM 提高了模糊支持向量机的分类性能,尤其是在 Fourclass,German 和 Heart 等数据集上的效果更加明显.在大多数情况下,RFSVM 的性能优于 $\nu$-SVM、RMSVM 和 SVM.这充分表明,对一个给定的数据集,同时考虑粗糙间隔和模糊隶属度,将会提高其分类性能.

其次,我们研究随机选择方法中训练数据比例的增加对分类结果的影响.为此,分别从数据集中随机选择 80% 与 90% 的数据做训练集,剩余的数据用于测试.实验结果如图 1 所示,其中横坐标的数字 1 ~ 7 分别表示不同的算法,即 1:RFSVM,2:RMSVM,3:$\nu$-SVM,4:FSVM100,5:SVM100,6:FSVM10 和 7:SVM10.可以看到,各个算法在大部分数据集上的分类正确率随着训练数据的增加而提高.

**表 1　不同方法在数据集上的平均正确率**

| Dataset | FRMSVM | RMSVM | $\nu$-SVM | FSVM-100 | SVM-100 | FSVM-10 | SVM-10 |
|---|---|---|---|---|---|---|---|
| Australian | 86.06 | 76.75 | 85.75 | 82.82 | 84.66 | 85.75 | 85.93 |
| Breast-cancer | 96.06 | 94.46 | 96.19 | 95.43 | 94.94 | 95.59 | 96.41 |
| Bupa | 71.68 | 59.18 | 68.97 | 66.96 | 70.63 | 55.59 | 66.78 |
| Cancer | 96.59 | 94.28 | 96.23 | 96.36 | 96.19 | 96.85 | 96.85 |
| Pima-Diabetes | 76.23 | 67.57 | 74.58 | 75.28 | 75.84 | 76.70 | 76.90 |
| Fourclass | 99.12 | 99.02 | 94.14 | 79.96 | 81.25 | 79.89 | 80.80 |
| German | 74.00 | 70.36 | 72.91 | 68.94 | 70.09 | 71.58 | 71.27 |
| Heart | 82.72 | 77.78 | 80.92 | 77.89 | 77.67 | 80.81 | 79.57 |
| Liver-disorders | 73.86 | 61.28 | 65.38 | 69.76 | 73.43 | 56.21 | 69.06 |
| Sonar | 87.73 | 88.46 | 88.89 | 87.73 | 88.46 | 88.89 | 88.46 |
| Splice | 79.45 | 76.88 | 76.88 | 75.85 | 76.88 | 71.15 | 76.88 |
| Wdbc | 97.93 | 97.56 | 95.96 | 97.34 | 97.08 | 97.88 | 97.77 |

为了验证提出的方法是否优于当前的算法,我们使用显著程度为 0.05 的双边 $t$ 检验.表 2 分别给出了 RFSVM 与其它算法在 12 个数据集上的 $t$ 检验结果,其中 Win,Tie 和 Loss 分别表示 RFSVM 算法明显优于、打平和劣于其它算法的数据集的个数.由表 2 可以看到,在这 12 个数据集中,本文提出的算法 RFSVM 优于其它

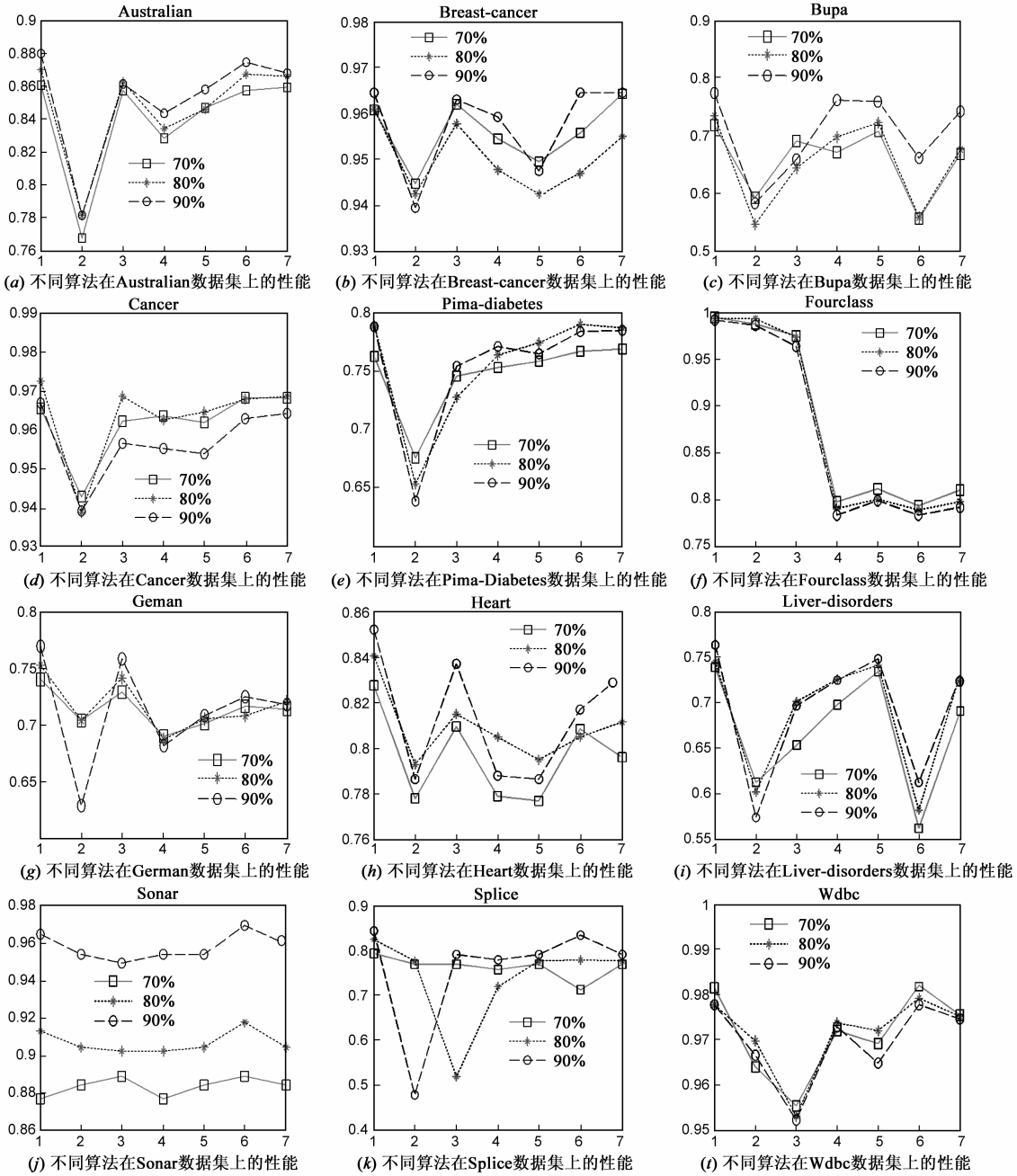
几种 SVM 算法.

为了表明提出的方法 RFSVM 在处理含有噪声或野点的数据集的优势,选择了 Bupa 与 Cancer 数据集进行了研究,实验结果如图 2($a$)和($b$)所示.

由图 3 知道,RFSVM 在处理噪声与野点的数据集优于其它 SVM 方法,并且随着噪声数据的增加,所有算法的性能有所下降.

**表 2 显著程度为 0.05 时双边 $t$ 检验的结果**

| 正确率(70%、80%和90%) | | Win | Tie | Loss |
|---|---|---|---|---|
| RFSVM | RMSVM | 10,8,8 | 2,4,4 | 0,0,0 |
| RFSVM | $\nu$-SVM | 5,5,5 | 7,7,7 | 0,0,0 |
| RFSVM | FSVM 100 | 8,6,6 | 4,6,6 | 0,0,0 |
| RFSVM | SVM 100 | 5,6,6 | 7,7,7 | 0,0,0 |
| RFSVM | FSVM 10 | 4,5,5 | 8,7,7 | 0,0,0 |
| RFSVM | SVM 10 | 5,3,3 | 7,9,9 | 0,0,0 |




($a$) 不同算法在Australian数据集上的性能 ($b$) 不同算法在Breast-cancer数据集上的性能 ($c$) 不同算法在Bupa数据集上的性能

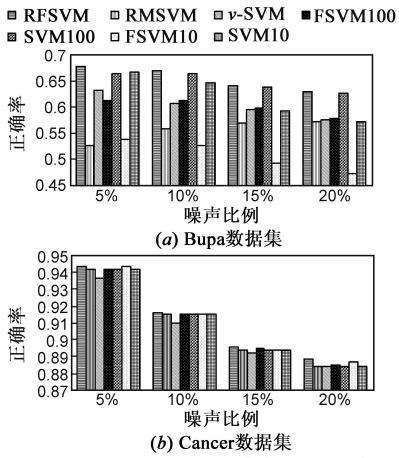
($d$) 不同算法在Cancer数据集上的性能 ($e$) 不同算法在Pima-Diabetes数据集上的性能 ($f$) 不同算法在Fourclass数据集上的性能

($g$) 不同算法在German数据集上的性能 ($h$) 不同算法在Heart数据集上的性能 ($i$) 不同算法在Liver-disorders数据集上的性能

($j$) 不同算法在Sonar数据集上的性能 ($k$) 不同算法在Splice数据集上的性能 ($l$) 不同算法在Wdbc数据集上的性能

**图1 选取不同比例数据的实验结果**

**(注: 纵坐标表示算法的性能,横坐标表示不同的算法,其中1:FRSVM,2:RMSVM,3:$\nu$-SVM,4:FSVM-100,5:SVM100,6:FSVM10,7:SVM10)**




(*a*) Bupa数据集

(*b*) Cancer数据集

图2　噪声与正确率间的关系

## 4　结论

本文采用粗糙集与模糊支持向量机相结合的方法,提出了基于粗糙间隔的模糊支持向量机,通过对训练样本赋予不同的模糊隶属度及引入上下间隔,以此来提高支持向量机的分类性能,更好地解决了含有噪声或野点数据对分类的影响,这种结合使得构造的最优决策函数更加合理;针对提出的算法在多个数据集上的实验结果表明,RFSVM 的分类性能优于其它支持向量机方法.另外,也要看到,尽管提出的 RFSVM 方法的性能优于其它方法,但本文只针对二分类问题进行了研究,在以后的工作中,需进一步研究多分类问题以及算法中涉及的参数取值,以便更好的用于实际问题之中.

### 参考文献


[1] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 2000.

[2] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20(3): 273 – 297.

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

[4] 周伟达,张莉,焦李成.线性规划支撑矢量机[J].电子学报,2001,29(11):1507 – 1511.
Zhou Weida, Zhang Li, Jiao Licheng. Linear programming support vector machines [J]. Acta Electronica Sinica, 2001, 29(11): 1507 – 1511. (in Chinese)

[5] 李昆仑,黄厚宽,田盛丰.模糊多类 SVM 模型[J].电子学报,2004,32(5):830 – 832.
Li Kunlun, Huang Houkuan, Tian Shengfeng. Fuzzy support vector machine for multi-class classification[J]. Acta Electronica Sinica, 2004, 32(5): 830 – 832. (in Chinese)

[6] Wang Y Q, Wang S Y, Lai K K. A support vector machine to evaluate credit risk [J]. IEEE Transactions on Fuzzy Systems, 2005, 13(6): 820 – 831.

[7] Zhang J H, Wang Y Y. A rough margin based support vector machine [J]. Information Sciences, 2008, 178(9): 2204 – 2214.

[8] Schölkopf B, Smola A J, Williamson R C, et al. New support vector algorithms [J]. Neural Computation, 2000, 12(5): 1207 – 1245.

[9] Xu Y T. A rough margin-based linear ν support vector regression[J]. Statistics and Probability Letters, 2012, 82(3): 528 – 534.

[10] Hsiao C C, Su S F, Chuang C C. A rough-based robust support vector regression network for function approximation[A]. Proceeding of International Conference on Fuzzy Systems[C]. Taipei, Taiwan: IEEE Press, 2011. 2814 – 2818.

[11] Chen D G, He Q, Wang X Z. FRSVMs: Fuzzy rough set based support vector machine[J]. Fuzzy Sets and Systems, 2010, 161(4): 596 – 607.

[12] Blake C L, Merz C J. UCI Repository of Machine Learning Databases[DB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

[13] King R D, Fenf C, Sutherland A. Statlog: comparison of classification algorithms on large real-world problems[J]. Applied Artificial Intelligence, 1995, 9(3): 289 – 333.

[14] Ho T K, Klernberg E M. Building projectable classifiers of arbitrary complexity[A]. Proceeding of the 13th International Conference on Pattern Recognition[C]. Vienna, Austria: IEEE Press, 1996. 880 – 885.


### 作者简介


**李　凯**　男,1963 年 9 月出生,河北满城人.2005 年毕业于北京交通大学计算机与信息技术学院,并获得工学博士学位.主要从事机器学习、数据挖掘、神经网络和模式识别等方面的研究工作.

E-mail: likai@hbu.edu.cn



**卢霄霞**　女,1984 年 5 月出生,河北行唐人.2012 年毕业于河北大学数学与计算机学院获工学硕士学位.主要从事机器学习和数据挖掘等方面的研究工作.