软间隔模糊粗糙支持向量机

鲁淑霞 忽丽莎 王熙照
(河北大学数学与计算机学院 保定 071002)

摘 要 分析了硬间隔模糊粗糙支持向量机(FRSVMs)的优点与不足。FRSVMs 通过修改硬间隔支持向量机(SVMs)的约束条件提高了泛化能力;FRSVMs 虽然将训练样例的条件属性与决策属性之间的不一致性考虑在内,但是在寻找最优超平面时仍然要求将训练集完全正确地分开,因此对噪音具有敏感性。针对 FRSVMs 的这个缺点,提出了软间隔模糊粗糙支持向量机(C-FRSVMs)。它使用高斯核函数作为模糊相似关系,将数据集中样例的条件属性与决策标签之间的不一致程度考虑在内;在训练寻找最优超平面的过程中允许存在错分点,并对原始最优化问题中训练样例的错分程度进行惩罚;既考虑了间隔最大,又考虑了训练误差最小,从而降低了对噪音的敏感性。实验表明:针对一些数据集,无论其是否存在异常点,C-FRSVMs 在测试精度上都可以同时优于硬间隔 SVMs、软间隔支持向量机(C-SVMs)和 FRSVMs,从而进一步提高了 FRSVMs 的泛化能力。

关键词 支持向量机,粗糙集,模糊粗糙集,模糊隶属度,模糊粗糙支持向量机
中图法分类号 TP181 **文献标识码** A

Fuzzy Rough Set Based Soft Margin Support Vector Machines

LU Shu-xia HU Li-sha WANG Xi-zhao
(College of Mathematics and Computer Science, Hebei University, Baoding 071002, China)

Abstract This paper analyzed the advantages and disadvantages of fuzzy rough set based support vector machines (FRSVMs). FRSVMs are generated by modifying constraints of hard margin support vector machines (SVMs) to get better generalization ability. Although having considered inconsistency between conditional attributes and decision attributes of training samples in datasets, FRSVMs construct the optimal hyperplane which must classify all the training samples correctly. So FRSVMs are sensitive to noises. Fuzzy rough set based soft margin support vector machines (C-FRSVMs) were proposed in this paper to overcome this shortcomings. C-FRSVMs use Gaussian kernel function as their fuzzy similarity relation, consider inconsistency between conditional attributes and decision labels of the samples in datasets, allow training samples to be misclassified during constructing the optimal hyperplane in the training process, punish the misclassification degrees of training samples in their original optimization problems. C-FRSVMs construct the optimal hyperplane by considering both maximal margin and minimal misclassification errors. So C-FRSVMs are less sensitive to noises than FRSVMs. Experimental results show that the proposed approach can obtain higher test accuracy compared with hard margin SVMs, soft margin support vector machines (C-SVMs) and FRSVMs. So, C-FRSVMs can get better generalization ability compared with FRSVMs.

Keywords Support vector machines, Rough set, Fuzzy rough set, Fuzzy membership, FRSVMs

1 引言

SVMs 主要用于解决分类问题。经典的 SVMs 在处理分类问题时,通过寻找条件属性与决策类标之间的关系生成决策函数,利用生成的决策函数预测新的测试样例。支持向量机作为一种在有限样本的情况下具有良好的学习性能和泛化性能的机器学习方法,已被成功地应用于人脸识别、数字水印、流量监测等领域[1,2]。然而 SVMs 没有考虑条件属性与决策类标之间的不一致性问题。粗糙集(RS)中明确地定义了一致性的概念,用依赖函数衡量条件属性与决策属性之间的一致程度,是一种处理模糊和不确定性知识的数学工具[3]。目前已有大量的学者提出 SVMs 与 RS 之间存在许多联系,并提出了许多将 SVMs 与 RS 相结合的模型。FSVM[4] 使用每个训练样例隶属于该样例的类标所在集合的隶属度对 C-SVMs 原始问题目标函数中该样例的错分程度进行惩罚,重新规划了 C-SVM;RMSVM[5] 将 υ-SVM 原始问题中间隔的定义粗糙化,对粗糙间隔中不同位置的训练样例对应的错分程度分配了不同的惩罚,重新规划了 υ-SVM;FRSVMs[6] 为

到稿日期:2010-09-22 返修日期:2010-12-24 本文受国家自然科学基金资助项目(60903088,60903089),河北省自然科学基金项目(F2010000323,F2011201063)资助。

鲁淑霞(1966-),女,博士,教授,主要研究方向为机器学习与计算智能、支持向量机,E-mail:mclsx@hbu.cn;忽丽莎(1986-),女,硕士,主要研究方向为支持向量机;王熙照(1963-),男,博士,教授,主要研究方向为模糊测度与模糊积分、模糊神经系统与遗传算法。

每个训练样例分配一个隶属度，在硬间隔 SVMs 的约束条件中通过每个样例的隶属度适当地放松约束条件，重新规划了硬间隔 SVMs。本文是对 FRSVMs 进行改进，通过放松 FRSVMs 的约束条件，在训练时允许出现错分点，从而降低了对噪音的敏感性。实验部分验证了 C-FRSVMs 的有效性。

2 基础知识

2.1 支持向量机

给定一个包含 ℓ 个训练样例的训练集 $T=\{(x_1,y_1),\cdots,(x_\ell,y_\ell)\}\in(X\times Y)^\ell$，其中 $x_i\in X=R^n$ 表示训练样例；$y_i\in Y=\{1,-1\}$，且 $i=1,\cdots,\ell$。如果训练集 T 线性不可分，通过核映射 Φ 将训练点映射到高维 Hibert 空间 H，使得训练集在高维空间 H 中线性可分。在高维空间中通过最大间隔的原则寻找最优超平面 $\omega\cdot\Phi(x)+b=0$，使训练点完全正确地分开，这就是硬间隔 SVMs；如果允许训练集 T 在高维空间 H 中是近似线性可分的，即允许有错分点，则通过综合考虑间隔最大化和错分程度最小化这两个原则寻找最优超平面 $\omega\cdot\Phi(x)+b=0$，使训练集大体上能够正确地分开，这就是 C-SVMs[7]。

2.2 粗糙集和模糊粗糙集

给定系统 $(U,A\cup D)$，其中：$U=\{x_1,\cdots,x_n\}$ 表示样例集合；A 中的属性称为条件属性；D 中的属性称为决策属性。在 $U\times U$ 上定义不可区分关系 $IND(A)$，$IND(A)$ 也是一个等价关系，设 $X\subset U$，定义 X 的下近似 $\underline{A}X$ 和上近似 $\overline{A}X$。位于 X 的下近似 $\underline{A}X$ 中的元素肯定属于 X；位于 X 的上近似 $\overline{A}X$ 中的元素可能属于也可能不属于 X。$(\underline{A}X,\overline{A}X)$ 就定义为 X 的粗糙集。定义 A 的 D 正域：

$$POS_A(D)=\bigcup_{X\in U/D}\underline{A}X$$

利用正域 $POS_A(D)$ 定义依赖函数 $\gamma_A(D)$，该函数用于测量条件属性 A 与决策属性 D 之间的一致度：如果 $\gamma_A(D)=1$，系统被称为是一致的；否则被称为是不一致的。$\gamma_A(D)$ 越大，一致性越强。保持 γ 不变的 A 的最小子集称为 A 的约简[8]。

在模糊粗糙集中，论域中两个样例之间的相似程度用模糊 T-相似关系描述。其中，T 是一个三角模。令 $F(U)$ 是 U 上的模糊幂集，则 $\forall X\in F(U)$，定义 X 关于模糊 $T-$相似关系 R 的上近似 $\overline{R}X$ 和下近似 $\underline{R}X$，X 的上近似 $\overline{R}X$ 和下近似 $\underline{R}X$ 都是模糊集。模糊集合 X 的模糊粗糙集定义为 $(\underline{R}X,\overline{R}X)$[9]。

2.3 模糊粗糙支持向量机

设条件属性集合 $A=\{a_1,\cdots,a_n\}$，决策属性集合 $D=\{1,-1\}$。对 $\forall t\in D_t$，D_t 表示类标是 t 的训练点组成的集合。FRSVMs 使用高斯核：

$$K_G(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$$

作为模糊 $T_{\cos}-$相似关系 R_G^n，即：

$$R_G^n(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$$

D_t 关于 R_G^n 的下近似和上近似分别为 $\underline{R_G^n}D_t$ 和 $\overline{R_G^n}D_t$。且：

$$\underline{R_G^n}D_t(x)=\inf_{u\in U}\upsilon_{T_{\cos}}(R_G^n(x,u),D_t(u))$$

$$\overline{R_G^n}D_t(x)=\sup_{u\in U}T_{\cos}(R_G^n(x,u),D_t(u))$$

当 $x\notin D_t$ 时，可以求得 $\underline{R_G^n}D_t(x)=0$；当 $x\in D_t$ 时，$s=\underline{R_G^n}D_t(x)$ 表示 x 属于集合 D_t 的隶属度。因此，FRSVMs 的原始最优化问题为[6]：

$$\min_{\omega\in H,b}r(\omega)=\frac{1}{2}\omega^2,s.t.\ y_i(\omega\cdot\Phi(x_i)+b)\geqslant s_i,i=1,\cdots,\ell \tag{1}$$

3 软间隔模糊粗糙支持向量机

FRSVMs[6] 考虑了条件属性与决策类标之间的一致性问题，是对硬间隔 SVMs 的改进，通过放松约束条件，使求出的最优超平面的间隔变大了，提高了泛化能力。然而 FRSVMs 在训练时要求所有的训练样例完全正确划分，因此对于线性不可分的训练集则通过使用高斯核映射将训练集映射到高维 Hibert 空间 H 使其线性可分。如果是由于噪声等因素引起的训练集不可分，则通过 FRSVMs 求出的超平面将有可能不是最优的，因此 FRSVMs 对噪音具有敏感性。为了克服以上缺陷，类似 C-SVMs 的构造方式，构造出 C-FRSVMs。

设训练集为 T，引进从输入空间 R^n 到一个高维 Hibert 空间 H 的变换：

$$\Phi:X\subset R^n\to X\subset H$$

$$x\mapsto\Phi(x)$$

C-FRSVMs 的原始最优化问题为：

$$\begin{aligned}&\min_{\omega\in H,b}r(\omega)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{\ell}\xi_i,\\&s.t.\ \ y_i(\omega\cdot\Phi(x_i)+b)\geqslant s_i-\xi_i,\\&\qquad\xi_i\geqslant 0;i=1,\cdots,\ell\end{aligned} \tag{2}$$

式中，s_i 表示 x_i 属于正类 $(y_i=1)$ 集合或负类 $(y_i=-1)$ 集合的隶属度。根据文献[6]，隶属度为：

$$s_i=R_G^nD_t(x_i)=\inf_{u\notin D_t}\sqrt{1-(R_G^n(x_i,u))^2}$$

设问题(2)的最优解为 ω^*，b^*。其相应的对偶问题为：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}y_iy_j\alpha_i\alpha_jK_G(x_i,x_j)-\sum_{j=1}^{\ell}s_j\alpha_j,\\&s.t.\ \sum_{i=1}^{\ell}\alpha_iy_i=0,\\&\qquad 0\leqslant\alpha_i\leqslant C;i=1,\cdots,\ell\end{aligned} \tag{3}$$

式中，K_G 表示高斯核。设对偶问题(3)的最优解为 $\alpha^*=(\alpha_1^*,\cdots,\alpha_\ell^*)$，则：

$$\omega^*=\sum_{i=1}^{\ell}y_i\alpha_i^*\Phi(x_i)$$

$$b^*=s_jy_j-\sum_{i=1}^{\ell}y_i\alpha_i^*K_G(x_i,x_j),0<\alpha_j^*<C$$

决策函数为：$f(x)=\mathrm{sgn}(\omega^*\cdot\Phi(x)+b^*)$。

4 实验结果

实验数据采用 UCI[10] 数据库中的 New_Thyroid，Wine，Parkinsons，Iris，Ionosphere，Tae 和 Sonar 七个数据集。因考虑的是两类的分类问题，故对于多类的数据集 New_Thyroid，Wine，Iris 和 Tae，通过合并类别的方式将它们转换成两类的数据集，即对于 New_Thyroid 和 Wine 两个数据集中的三类数据，均把第一类作为正类，第二类和第三类合并作为负类；对于 Iris 中的三类数据，把第一类和第三类合并作为正类，第二类作为负类；对于 Tae 中的三类数据，把第一类和第二类合并作为正类，第三类作为负类。数据集的详细信息如表 1 所列。

表 1　数据集信息

数据集	样例个数	正例个数	负例个数	属性个数	类别个数
New_Thyroid	215	150	65	5	2
Wine	178	59	119	13	2
parkinsons	195	147	48	22	2
Iris	150	100	50	4	2
Ionosphere	351	225	126	34	2
Tae	151	99	52	5	2
Sonar	208	97	111	60	2

整个实验是在个人计算机(配置:英特尔酷睿 i3-370m,2.40MHZ,320-GB 硬盘,2GB 内存)上使用 MATLAB SVM-KM 工具箱实现的。其中,硬间隔 SVMs,C-SVMs,FRSVMs 和 C-FRSVMs 这 4 种分类器在训练时均使用高斯核。对于每个数据集,从 1～10 这 10 个整数中找出使硬间隔 SVMs 的测试精度最高的高斯核参数 σ,将它作为 4 种分类器关于该数据集进行分类时的核参数;惩罚参数 C 取固定值 100。

除了考虑这 4 种分类器在上述 7 个数据集上的分类预测情况外,还考虑了当上述数据集包含异常点时,这 4 种分类器对含有异常点的数据集进行分类预测的情况。含有异常点的数据集按照以下两种方式构造:

①从数据集中随机选取一些样例,改变这些样例的类标再放回数据集中;

②所有的训练集进行一次硬间隔 SVMs 训练,得到一个最优超平面。从数据集中选取一些位于超平面附近且被正确分类的样例,改变这些样例的类标再放回数据集中。

使用这 4 种分类器分别对不含有异常、含有 10%,20%,30%的①类异常点和含有 10%,20%,30%的②类异常点的数据集进行分类预测,实验过程采用 10 次交叉验证。将获得的训练精度值、测试精度值和测试精度的方差值进行了比较,这里的训练精度值和测试精度值均是 10 次交叉验证结果的平均值。实验结果如表 2－表 8 所列。

表 2　4 种分类器在 New_Thyroid 数据集上的分类预测结果(σ=2)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.9714±0.0033	1	0.981±0.0019	1	0.9762±0.0015	1	0.9857±4.76E-044
①10%	1	0.8143±0.0038	1	0.8286±0.0046	1	0.8429±0.0086	1	0.8619±0.0016
①20%	1	0.6429±0.0119	1	0.6619±0.0061	1	0.6905±0.0102	1	0.719±0.0097
①30%	1	0.5857±0.0114	1	0.6286±0.0076	1	0.6143±0.0199	1	0.6524±0.0073
②10%	1	0.7952±0.0077	1	0.819±0.0076	1	0.8238±0.0041	1	0.8381±0.0028
②20%	1	0.681±0.0073	1	0.7048±0.0058	1	0.7381±0.0069	1	0.7429±0.0124
②30%	1	0.6667±0.0077	1	0.6905±0.0078	1	0.6857±0.006	1	0.7±0.0037

表 3　4 种分类器在 Wine 数据集上的分类预测结果(σ=9)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.8588±0.0057	1	0.8765±0.0114	1	0.8706±0.0075	1	0.8882±0.0022
①10%	1	0.7235±0.007	1	0.7471±0.0063	1	0.7588±0.0142	1	0.8±0.0042
①20%	1	0.6294±0.0118	1	0.6588±0.0089	1	0.6529±0.0148	1	0.6941±0.0043
①30%	1	0.5706±0.0166	1	0.6294±0.0229	1	0.6118±0.0216	1	0.6471±0.0166
②10%	1	0.8294±0.0052	1	0.8471±0.0112	1	0.8412±0.0083	1	0.8647±0.0026
②20%	1	0.8706±0.0151	1	0.8882±0.0052	1	0.8824±0.0048	1	0.9±0.0014
②30%	1	0.9765±0.0015	1	0.9765±0.0015	1	0.9765±0.0015	1	0.9824±0.0014

表 4　4 种分类器在 Parkinsons 数据集上的分类预测结果(σ=10)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.9±0.008	0.9988	0.9053±0.0016	1	0.9158±0.0034	1	0.9263±0.0023
①10%	1	0.7684±0.009	0.9912	0.8±0.0071	1	0.7842±0.0113	1	0.8158±0.0057
①20%	1	0.6895±0.0025	0.9883	0.7432±0.0129	1	0.6947±0.0099	1	0.7526±0.0122
①30%	1	0.5895±0.0049	0.9895	0.6±0.0117	1	0.6211±0.0165	1	0.6316±0.0116
②10%	1	0.7632±0.004	0.9947	0.7684±0.0073	1	0.7842±0.0091	0.9994	0.8±0.0027
②20%	1	0.7316±0.0135	0.9947	0.7474±0.0082	1	0.7526±0.0089	1	0.7632±0.0062
②30%	1	0.7263±0.0147	0.9942	0.7368±0.0116	1	0.7316±0.0115	0.9994	0.7421±0.0063

表 5　4 种分类器在 Iris 数据集上的分类预测结果(σ=1)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.9333±0.0027	0.9889	0.9533±0.0045	1	0.9533±0.0036	0.9948	0.9667±0.0011
①10%	1	0.7733±0.0037	0.9281	0.8533±0.0087	1	0.8267±0.0082	0.9622	0.8667±0.0044
①20%	0.9933	0.6933±0.0091	0.8681	0.7533±0.0072	0.9926	0.72±0.0132	0.9296	0.76±0.012
①30%	0.9933	0.6067±0.0072	0.8519	0.6667±0.0107	1	0.64±0.0142	0.8733	0.68±0.0047
②10%	0.9933	0.8±0.006	0.903	0.82±0.0072	0.9874	0.8067±0.0093	0.9267	0.8267±0.0055
②20%	0.9941	0.7667±0.0091	0.8978	0.7867±0.0069	0.9904	0.7733±0.0117	0.9156	0.7933±0.0054
②30%	1	0.8667±0.0132	0.9563	0.9067±0.0037	1	0.86±0.0031	0.9689	0.9133±0.002

表 6　4 种分类器在 Ionosphere 数据集上的分类预测结果($\sigma=1$)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.9486±0.0016	1	0.9514±9.88E-04	1	0.9514±0.0013	1	0.9543±3.35E-04
①10%	1	0.7971±0.003	0.9933	0.8029±0.0073	1	0.8143±0.0113	0.9952	0.82±0.0011
①20%	1	0.6714±0.0066	0.9889	0.7086±0.005	1	0.7029±0.0115	0.9946	0.7286±0.0036
①30%	1	0.5486±0.006	0.9924	0.6057±0.0018	1	0.62±0.0028	0.9952	0.6343±0.0055
②10%	1	0.8371±0.0021	1	0.8514±0.0047	1	0.8514±0.0032	1	0.86±0.0016
②20%	1	0.7829±0.0018	1	0.7857±0.0046	1	0.7829±0.0022	1	0.7943±0.0024
②30%	0.9971	0.6914±0.0044	0.9971	0.7029±0.0067	0.9975	0.7114±0.0042	0.9975	0.72±0.0015

表 7　4 种分类器在 Tae 数据集上的分类预测结果($\sigma=1$)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	0.9815	0.8067±0.0128	0.9815	0.8133±0.0043	0.9807	0.8133±0.0123	0.9807	0.8267±0.0055
①10%	0.9407	0.72±0.0087	0.9452	0.7467±0.006	0.9393	0.7467±0.0128	0.963	0.8±0.0053
①20%	0.9267	0.6667±0.0267	0.9215	0.68±0.006	0.9207	0.68±0.0114	0.92	0.6867±0.0027
①30%	0.8896	0.5667±0.0082	0.903	0.62±0.0223	0.8978	0.6±0.0151	0.8852	0.62±0.0114
②10%	0.94	0.6933±0.0162	0.9393	0.7267±0.0093	0.9393	0.7267±0.0244	0.937	0.74±0.004
②20%	0.897	0.6133±0.0096	0.8978	0.64±0.0108	0.8978	0.6267±0.0028	0.8941	0.6467±0.0064
②30%	0.8578	0.5133±0.0116	0.8556	0.54±0.0217	0.8548	0.54±0.0137	0.8541	0.5533±0.0036

表 8　4 种分类器在 Sonar 数据集上的分类预测结果($\sigma=1$)

异常点	SVM		C-SVM		FRSVMs		C-FRSVMs	
	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度	训练精度	测试精度
无异常点	1	0.895±0.0037	1	0.9±0.004	1	0.905±0.0102	1	0.91±0.0014
①10%	1	0.795±0.0047	1	0.815±0.012	1	0.79±0.0051	1	0.825±0.0011
①20%	1	0.67±0.0059	1	0.69±0.0074	1	0.675±0.006	1	0.69±0.0024
①30%	1	0.565±0.0195	1	0.59±0.0064	1	0.57±0.0081	1	0.605±0.0147
②10%	1	0.86±0.0029	1	0.865±0.006	1	0.86±0.0119	1	0.87±0.0056
②20%	1	0.905±0.0032	1	0.915±0.008	1	0.905±0.0047	1	0.91±0.0024
②30%	1	0.945±0.0032	1	0.96±0.009	1	0.95±0.002	1	0.96±9.00E-04

根据表 2,表 4,表 6,表 7 中的测试精度值的比较可以看出:随着异常点个数的增加,上述 4 个分类器的测试精度均逐渐降低。C-FRSVMs 关于这 4 个数据集基本上都达到了最高的测试精度。

根据表 3,表 5,表 8 中的测试精度值的比较可以看出,随着①类异常点个数的增加,上述 4 个分类器的测试精度均逐渐降低;但随着②类异常点个数的增加,上述 4 个分类器的测试精度均出现上升的趋势。导致后一种情况出现的原因可能是此时的第②类异常点都位于超平面附近,是那些容易被分错的样例。原来的数据集由于异常点的加入而变得更容易被分开,使错分点变少了。C-FRSVMs 在包含两类异常点的数据集上大都可以获得最高的测试精度;当 Sonar 数据集中分别包含 20%的①类异常点和 30%的②类异常点时,C-SVMs 也可以获得最高的测试精度;当 Sonar 数据集中包含 20%的②类异常点时,C-SVMs 达到了最高的测试精度,C-FRSVMs 的测试精度略低于 C-SVMs。

下面分析 4 种分类器在 7 个数据集上获得的不同测试精度的差异情况。通过比较 10 次交叉验证在各个数据集上测试精度的方差值可以看出:C-FRSVMs 关于这 7 个数据集基本上都可以获得最小的方差。特别地,当 Sonar 数据集中分别包含 20%的①类异常点、30%的②类异常点和 20%的②类异常点时,尽管 C-SVMs 的测试精度略高于 C-FRSVMs,但它关于测试精度的方差却低于 C-FRSVMs,这意味着 C-FRSVMs 的测试结果比 C-SVMs 的结果更稳定。

结束语　本文基于 FRSVMs 的不足提出了 C-FRSVMs。FRSVMs 在寻找最优超平面时要求训练样例必须被完全正确地分开,因此对噪音具有敏感性。C-FRSVMs 通过软化对间隔的要求放松约束条件,在训练的过程中允许有错分点。因此,C-FRSVMs 对最优超平面选取的过程减少了对数据集中噪音的敏感性。实验结果表明,与硬间隔 SVMs、C-SVMs、FRSVMs 相比,C-FRSVMs 一般能获得更好的测试精度,具有更好的预测能力。

参 考 文 献

[1] 李春花,凌贺飞,卢正鼎. 基于支持向量机的自适应图像水印技术[J]. 计算机研究与发展,2007,44(8):1399-1405

[2] 吴敏,王汝传. 一种基于损失函数的 SVM 算法在 P2P 流量监测中的作用[J]. 计算机科学,2009,36(12):76-80

[3] Pawlak Z. Rough Set International[J]. Information Sciences, 1982,11(5):341-356

[4] Lin Chun-fu, Wang Sheng-de. Fuzzy Support Vector Machine [J]. IEEE Trans Neural Networks,2002,13(2):464-471

[5] Zhang Jun-hua, Wang Yuan-yuan. A rough magin based support vector machine[J]. Information Sciences, 2008, 178(9): 2204-2214

[6] Chen De-gang, He Qiang, Wang Xi-zhao. FRSVMs: Fuzzy rough set based support machines[J]. Fuzzy Sets and Systems, 2010, 161(4):596-607

[7] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2004

[8] Pawlak Z. Rough Sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991

[9] Radzikowska A M, Kerre E E. A comparative study of fuzzy rough sets[J]. Fuzzy Sets and Systems, 2002, 126(2): 137-155

[10] Murphy P M, Aha D W. UCI machine learning repository[DB/OL]. http://www.ics.uci.edu/mlearn/ML Repository, 2010-11-6