基于不平衡数据分类的一种平衡模糊支持向量机

秦传东[1] 刘三阳[2] 张市芳[1]

(西安电子科技大学计算机学院 西安 710071)[1] (西安电子科技大学理学院 西安 710071)[2]

摘 要 鉴于不平衡数据集中类不平衡比较大的分类问题,利用样本点的特性建立类不平衡调节因子和模糊隶属度,提出了平衡模糊支持向量机。首先计算样本协方差矩阵,求得类不平衡调节因子,然后计算各样本点的模糊隶属度,得到各样本对分类超平面的贡献率。类平衡调节因子和模糊隶属度同时对分类器的误差项产生影响。结果表明,这种平衡模糊支持向量机对类不平衡比较大的分类问题具有很好的分类效果。

关键词 支持向量数据域描述,模糊隶属度,模糊支持向量机,平衡模糊支持向量机,不平衡因子

中图法分类号 TP181 **文献标识码** A

Balanced Fuzzy Support Vector Machine Based on Imbalanced Data Set

QIN Chuan-dong[1] LIU San-yang[2] ZHANG Shi-fang[1]

(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)[1]

(Dept. of Mathematic Science, Xidian University, Xi'an 710071, China)[2]

Abstract In view of the classification of imbalance data set with the larger imbalanced ratio of class, a balanced fuzzy support vector machine(BFSVM) was proposed, making use of the imbalance adjustment factor and the fuzzy membership based on the features of sample points. Firstly, it computes the sample covariance matrix and gets the imbalance adjustment factor, then computes the fuzzy membership of every sample and gets the contribution rate of every sample. Fuzzy membership and imbalance adjustment affect the sample error of classifier at the same time. The experiment results prove that the algorithm has a good effect on the larger imbalanced ratio.

Keywords Support vector data description, Fuzzy membership, Fuzzy support vector machine, Balanced fuzzy support vector machine, Imbalanced factor

1 引言

标准支持向量机(Support Vector Machines,简称 C-SVM)是由 Vapnik 等[1,2]创立的一种建立在统计学习理论和结构风险最小化原理基础上的优化算法,能较好地解决小样本、高维数、非线性等问题。当人们把这种 C-SVM 应用到不平衡数据集分析时,发现 C-SVM 对少数样本(Minority,称为正类,简称 N^+)分类和预测的准确率远低于对多数(Majority,称为负类,简称 N^-)样本的分类准确率,即算法多具有偏向于多类的特性。这使 C-SVM 算法对不平衡数据集的分类性能大大下降[3-5]。针对这些分类问题,很多研究从不同的角度来研究不平衡数据的特性,以便提高分类性能。有研究者[9]提出对正类(N^+)采用上采样(Oversampling)和对负类(N^-)采用下采样(Undersampling)等方法来弥补与负类的差距,以达到平衡作用。但这些方法实际上是把相邻的边界点去掉,从而会失去一些有用的信息,随机性也难以保证,结果自然也不可靠。文献[7]提出一种不平衡支持向量机的调整方法,它只考虑到正负类数据的不平衡性,而没有考虑到样本点的疏密对分类超平面的影响。文献[8]提出一种近似支持向量机算法,以处理不平衡问题中的正、负类样本所赋予的惩罚因子。文献[9,10]提出对不平衡数据集正、负类松弛变量给出不同的惩罚因子的方法(Different Error Costs, DEC)。据此原理,很多学者将正类或负类数据量的多少或者数据的不平衡比直接作为平衡因子来调整数据集的不平衡性,这有时是比用标准的支持向量机的效果好点,但它忽略了样本点的疏密对分类超平面的影响,而且对不平衡比过大的数据集也不适用。综合考虑不平衡数据集的不平衡比、数据集中的样本的疏密、离分类超平面的远近等因素,本文提出一种平衡模糊支持向量机(Balanced Fuzzy Support Vector Machine, BFSVM),引入模糊隶属度和类不平衡调节因子,考虑每个样本点对超平面的贡献率和整体样本的不平衡对分类器的影响,从点到面全面考虑了不平衡数据集中的样本点对分类超平面的影响。

2 支持向量机简介

给定样本集$\{(x_i, y_i)\}$,$x_i \in R^n$,$y_i \in \{-1, 1\}$,其中 $i=1, \cdots, m$。引入适当的映射 $\varphi: x_i \rightarrow \varphi(x_i)$,将样本 x_i 映射到高维特征空间中。选取适当的核函数,使得 $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$。引入松弛变量 $\xi_1, \xi_2, \xi_3, \cdots, \xi_m$ 及惩罚因子 C,在最小化经验风险的原则下,利用 Tikhonov 正则化框架理论[1],得

到稿日期:2011-07-01 返修日期:2011-11-03 本文受国家自然科学基金(60974082)资助。

秦传东(1976—),男,博士生,主要研究方向为机器学习、模式识别。

到标准支持向量机(简称 C-SVM)的一般模型：

$$\min_{\omega\in R^n,\xi_i\in R}\frac{1}{2}\omega^{\mathrm{T}}\omega+C\sum_{i=1}^{m}\xi_i$$
$$\text{s. t. } y_i(\omega\cdot\varphi(x_i)+b)\geqslant 1-\xi_i \tag{1}$$
$$\xi_i\geqslant 0, i=1,2,3,\cdots,m$$

其对偶优化问题[11]为

$$\min\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j k(x_i,x_j)-\sum_{i=1}^{m}\alpha_i$$
$$\text{s. t. }\sum_{i=1}^{m}y_i\alpha_i=0 \tag{2}$$
$$0\leqslant\alpha_i\leqslant C, i=1,2,3,\cdots,m$$

问题(2)的最优解是 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)^{\mathrm{T}}$。

求得超平面的法向量为

$$\omega^*=\sum_{i=1}^{m}\alpha_i^* y_i\varphi(x_i) \tag{3}$$

选取某个 $0<\alpha_j^*<C$ 所对应的 x_j,y_j，求得

$$b^*=y_j-\sum_{i=1}^{m}y_i\alpha_i k(x_i,x_j) \tag{4}$$

由此得到决策函数为

$$f(x)=\sum_{i=1}^{m}y_i\alpha_i k(x_i,x)+b^* \tag{5}$$

3 平衡调节

3.1 不平衡调节因子

对不平衡数据而言，正则化框架理论下求得该超平面具有偏移性，这与两类样本数和样本分散程度有关。考虑样本数的差异和样本的分散程度对分离超平面的影响十分必要。又知协方差矩阵用于计算不同维度数据的样本点偏离其均值的程度。本文考虑样本点沿分离超平面的法向量 ω 方向的协方差矩阵，用协方差矩阵的投影标准差来表示样本的分散程度。投影后正、负类标准差[12]分别为

$$s^+=\sqrt{\omega^{\mathrm{T}}\Sigma^+\omega}, s^-=\sqrt{\omega^{\mathrm{T}}\Sigma^-\omega}$$

式中，$\Sigma^+\Sigma^-$ 分别是正、负类的协方差矩阵。对于给定的数据样本集，使用 C-SVM 进行初步训练，计算投影值为

$$\omega_1\varphi(x_j)=\sum_{i=1}^{m}\alpha_i^* y_i k(x_i,x_j), j=1,2,\cdots,m \tag{6}$$

然后分别计算 s^+ 和 s^-，即

$$s^+=\sqrt{\frac{1}{n^+-1}}\cdot\sqrt{\sum_{j=1}^{n^+}[\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j)-\frac{1}{n^+}\sum_{j=1}^{n^+}\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j)]^2}$$

$$s^-=\sqrt{\frac{1}{n^--1}}\cdot\sqrt{\sum_{j=1}^{n^-}[\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j)-\frac{1}{n^-}\sum_{j=1}^{n^-}\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j)]^2}$$

分析表明，使用 C-SVM 时，分离超平面会向样本数少的类即正类偏移。为了纠正偏移性，采用不同惩罚因子 C^+,C^- 的比例 $\frac{C^+}{C^-}\propto\frac{n^-}{n^+}$。又通过以上分析知 $\frac{C^+}{C^-}\propto\frac{s^+}{s^-}$，则 $\frac{C^+}{C^-}\propto\frac{s^+}{s^-}\cdot\frac{n^-}{n^+}$。从而惩罚因子可定义为 $C^+=C\frac{s^+}{n^+}, C^-=C\frac{s^-}{n^-}$，$C$ 是正常数。

3.2 模糊隶属度

样本的分散程度对分类超平面的影响不仅要考虑样本的协方差矩阵，而且要考虑每个样本点对超平面的贡献率，即模糊。考虑到这一点，设样本点离各自类中的最大距离为

$$\begin{cases}x_{\max}=\max\|x_i-\overline{x_+}\|, x_i\in A^+\\ x_{\max}=\max\|x_i-\overline{x_-}\|, x_i\in A^-\end{cases} \tag{7}$$

则定义模糊隶属度

$$m_i=\begin{cases}1-\dfrac{x_i-\overline{x_+}}{x_{\max}+\delta}, & x_i\in A^+\\ 1-\dfrac{x_i-\overline{x_-}}{x_{\max}+\delta}, & x_i\in A^-\end{cases} \tag{8}$$

式中，δ 是个任意小的正数，A^+,A^- 表示正类样本集和负类样本集。从定义可以看出，处于类中心的样本点被赋予比较高的隶属度，而离类中心比较远的野点或者噪声被赋予较小的隶属度，这样可以减少野点或噪声点对实验的影响。

4 平衡模糊支持向量机

4.1 平衡模糊支持向量机

求得数据集的不平衡调节因子和样本的隶属度是 m_i 之后，由 C-SVM 的基本思想，对正类和负类的松弛变量运用不平衡调节因子和样本的模糊隶属度，得到下面的优化问题：

$$\min_{\omega,y}\frac{1}{2}\omega^{\mathrm{T}}\omega+C^+\sum_{\{i|y=+1\}}^{m}m_i\xi_i+C^-\sum_{\{i|y=-1\}}^{m}m_i\xi_i$$
$$\text{s. t. } y_i(\omega\cdot\varphi(x_i)+b)\geqslant 1-\varepsilon_i \tag{9}$$
$$\varepsilon_i>0, i=1,\cdots,m$$

式中，$C^+=C\frac{s^+}{n^+}, C^-=C\frac{s^-}{n^-}$，而 C 是一个正的常数。利用 Lagrange 函数运用 KKT 条件，求得其对偶优化问题为

$$\min_{\alpha\in R^n}\frac{1}{2}\sum_{i}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i\cdot x_j)-\sum_{i=1}^{m}\alpha_i$$
$$\text{s. t. }\sum_{i=1}^{m}\alpha_i y_i=0$$
$$0\leqslant\alpha_i\leqslant Cm_i\frac{s^+}{n^+}, y_i=1, i=1,2,\cdots,n^+ \tag{10}$$
$$0\leqslant\alpha_i\leqslant Cm_i\frac{s^-}{n^-}, y_i=-1, i=1,2,\cdots,n^-$$

由上面的优化问题(10)求得最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)^{\mathrm{T}}$，进而求解偏值及决策函数

$$\begin{cases}b^*=y_j-\sum_{i=1}^{m}y_i\alpha_i k(x_i,x_j)\\ f(x)=\mathrm{sgn}(\sum_{i=1}^{m}\alpha_i^* y_i k(x_i,x)+b^*)\end{cases} \tag{11}$$

4.2 算法步骤

①选择适当的高斯核函数 $K(x,x^{\mathrm{T}})$ 和惩罚参数 C，利用 C-SVM 得到最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)^{\mathrm{T}}$ 和超平面的法向量的投影值 $\omega_1\varphi(x_j)=\sum_{i=1}^{m}\alpha_i^* y_i k(x_i,x_j), j=1,2,\cdots,m$。

②将①中的高斯核函数和上面求得的投影值代入 s^+,s^- 中，求得正类和负类的投影标准差 s^+ 和 s^-，结合正类数 n^+、负类数 n^-，得到不平衡调节因子 C^+ 和 C^-。

③ 计算样本点的模糊隶属度 m_i。

④ 利用 C^+ 和 C^- 及 m_i，构造并求解最优化问题(10)，得到其最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)^{\mathrm{T}}$。选择一个 α^* 的分量 $\alpha_j^*\leqslant Cm_j\cdot\frac{s^+}{n^+}$ 或 $\alpha_j^*\leqslant Cm_j\cdot\frac{s^-}{n^-}$，由此有分类超平面的偏置，进而判别函数(11)。

5 实验与结果分析

为了检测所提算法的性能，从 UCI 机器学习库中取出 5

个基准分类问题的数据。数据集[13]的各个属性如表1所列。

表1 5个不平衡数据集的属性

数据集	正类	负类	总数	不平衡比 R
Abanole	103	4074	4177	39.55
Yeast	51	1433	1484	28.10
Satimage	626	5809	6435	9.28
Haberman	81	225	306	2.78
Pimaindians	268	500	768	1.87

试验时，先将数据归一化，然后选择70%的数据进行训练，用剩下的30%进行测试，做算法比较。实验中的核函数采用广泛应用的径向基高斯核函数 $k(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$。参数 γ 和惩罚参数 C 采用网格搜索的方法对其择优选择，选用最好的校验正确率所对应的参数训练整个训练集，并在测试集上进行测试，选用测试集上的分类评价标准作为算法的分类标准。实验是在 inte(R) Core(TM) 2 CPU 6320 @1.86G，1G 内存的 PC 机上安装的 Matlab R2009 软件上实现的。SVM 是利用网站 http://theoval.sys.uea.ac.uk/svm/toolbox 提供的 Matlab SV 软件工具包来实现的。

对高不平衡数据集来说，已经不宜采用分类准确率来评价试验效果，而是采用众多不平衡数据学习中的 Se(Sensitivity)，Sp(Specificity)和 Gm(Gmean)来评价[4]，其定义为

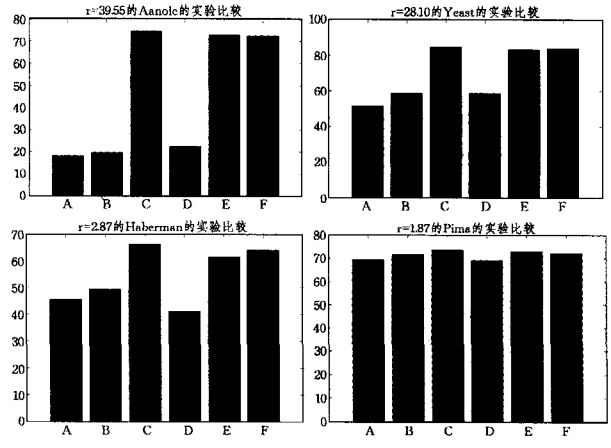
$$Se=\frac{TP}{TP+FN};Sp=\frac{TN}{TN+FP};Gm=\sqrt{Se\cdot Sp}$$

式中，TP(True Positive)、FN(False Negative)、TN(True Negative)以及 FP(False Positive)依次代表分类正确的正类样本、假的负类样本、正确的负类样本以及假的正类样本的数目，且有 $TP+FN=N^+$、$TN+FP=N^-$ 以及 $N^++N^-=N$。Se 和 Sp 分别度量了分类器能正确分辨正类和负类样本的能力，由定义知 Se 和 Sp 越大越好。但在许多情况下，一个具有高 Se 的分类器不一定具有高的 Sp。因此，同时利用它们来评价分类器的性能是比较困难的，故引入其几何均值 $Gm=\sqrt{Se\cdot Sp}$ 来共同评价分类器优劣。Gm 越大，分类效果越好。实验先将 BFSVM 同 DEC、FSVM 比较，然后把 BFSVM 算法同 C-SVM，Undersampling[9]，Oversampling[9]进行比较。

表2 BPSVM算法与DEC，FSVM算法的比较结果(DEC算法采用不平衡比作为调节因子)

数据集	算法	Se	Sp	Gm
Abanole	DEC	3.23	99.56	17.93
	FSVM	3.82	99.44	19.49
	BFSVM	83.26	66.86	**74.61**
Yeast	DEC	27.30	98.02	51.73
	FSVM	35.09	98.40	58.76
	BFSVM	83.24	86.82	**85.01**
Satimage	DEC	67.52	97.30	81.05
	FSVM	70.20	97.08	82.55
	BFSVM	89.90	91.60	**90.75**
Haber man	DEC	22.30	93.56	45.68
	FSVM	27.21	89.78	49.30
	BFSVM	58.20	75.12	**66.12**
Pima-Indians	DEC	60.67	78.91	69.19
	FSVM	66.79	76.63	71.54
	BFSVM	70.62	76.21	**73.36**

从表2和图1来看，BFSVM算法比FSVM，DEC算法具有更强的优越性，特别是在数据集的不平衡比例比较大时。例如 Abanole 的负、正类不平衡比是 39.55，这时如用算法 BFSVM，求得的 Gm 值比 DEC 和 FSVM 的具有较大的改善。随着类不平衡比降低到 Pima-indians 的 1.87 时，算法 BFSVM 的优越性也逐渐降低。Gm 的优越性不是很显著，但还是比单一用不平衡比作调节因子的 DEC 算法和仅仅用模糊隶属度的 FSVM 算法要好，这主要是因为即使数据集正、负类数据相当，趋于平衡，调节平衡因子和样本点的模糊隶属度也能提高算法的精确性。这一点从表2的黑体数字可以看出。



其中 A,B,C,D,E,F 分别代表：DEC，FSVM，BFSVM，C-SVM，Undersampling，Oversampling 方法

图1 4种不同不平衡比数据在6种方法下实验效果 Gm 值对比图

从表3的实验数据和图1来看，BFSVM算法得到的分类器的性能比传统的标准支持向量机要好，不管数据集是平衡的还是不平衡的。但是与 Undersampling 和 Oversampling 两种算法相比较，效果不是那么明显。它们的 Gm 值相近，这说明算法性能的好坏与数据集的不平衡程度、数据集中样本点的分布特点有很大的关系。对高不平衡比的数据集来说，BFSVM 得到的分类器性能要优于低不平衡比的数据集所得的分类器性能。当然 BFSVM 在求解不平衡因子时需要对数据集进行初次测试，求得不平衡调节因子。这需要消耗近一半的时间，也是 BFSVM 算法的不利之处。

表3 BFSVM算法同C-SVM，Undersampling、Oversampling 3种算法的实验结果比较

数据集	算法	Se	Sp	Gm
Abanole	C-SVM	5.06	98.2	22.28
	Under.	79.12	67.66	73.17
	Over.	74.11	71.38	72.73
	BFSVM	83.26	66.86	74.61
Yeast	C-SVM	35.09	98.12	58.68
	Under.	80.89	85.77	83.29
	Over.	81.32	86.70	84.00
	BFSVM	83.24	86.82	**85.01**
Satimage	C-SVM	68.79	96.88	81.64
	Under.	91.84	86.46	89.11
	Over.	88.21	89.09	88.64
	BFSVM	89.90	91.60	**90.75**
Haber-man	C-SVM	19.56	89.76	40.90
	Under.	51.49	73.97	61.71
	Over.	53.16	76.89	63.93
	BFSVM	58.20	75.12	**66.12**
Pimaindians	C-SVM	52.82	89.93	68.92
	Under.	70.56	74.80	72.65
	Over.	69.31	75.12	72.20
	BFSVM	70.62	76.21	**73.36**

(下转第212页)

表 1　文本的类分布

类别	文件数	类别	文件数
游戏	22843	旅游	18471
经济	40115	文艺	14248
科技	53126	时政_国际	59130
房产	19573	时政_国内	119695
汽车	21745	教育	24405
体育	96120	生活男女	19382
娱乐	23905	时政_社会	42559
时政_军事	21743	总计	597060

本文以“经济”为例计算了该分类的准确度，如表 2 所列。

表 2　以“经济”为例的准确率

领域	正确词数	抽取到的总词数	准确率
经济	962	1000	96.2%
	1916	2000	95.8%
	2870	3000	95.6%
	3814	4000	95.3%
	4737	5000	94.7%

从表 2 中可以看出，本文提出的算法文本分类准确率都高达 95%以上，为了更进一步说明本文算法比传统的算法优越，我们比较了常见的支持向量机文本分类方法、遗传算法、聚类算法和传统的最大熵文本分类算法，比较结果如表 3 所列。

表 3　不同算法的性能比较

	时间(s)	召回率(%)	正确率(%)
聚类算法	47.52	82.32	90.43
遗传算法	26.31	89.12	93.22
支持向量机方法	24.34	90.13	93.32
传统的最大熵算法	20.12	93.21	96.75

从表 3 中可以看出本文所提算法无论在时间上还是正确率都具有一定的优越性。

结束语　将许多的文本资料根据属性或者内容划分到相对应的类别中的过程是文本分类，这在日常生活中比较普遍，但是传统的文本算法都存在着特征词不清楚、分成的类别有重叠的结果的缺陷。针对这种效率不高的文本算法，本文提出了一种新算法，其采用了最大熵算法和均值聚类算法，并将这两种算法相结合，可以降低算法的复杂度。较之传统算法而言，它可以在较短的时间内获得每一个特征子集，准确率得到了很大的提高。

参 考 文 献

[1] Huang Zhe-xue, Michael K N, Rong hong-qiang, et al. Automated Variable Weighting in k-Means Type Clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[2] 刘斌，黄铁军，程军，等. 一种新的基于统计的自动文本分类方法[J]. 中文信息学报，2002，16(6)：18-24

[3] 朱明，王俊普. 一种最优特征集的选择算法[J]. 计算机研究与发展，1998，9(35)：803-805

[4] 胡佳妮，徐蔚然，郭军，等. 中文文本分类中的特征选择算法研究[J]. 光通信研究，2005(3)：44-46

[5] 潘有能. 一个自动分词分类系统的实现[J]. 情报学报，2002，21(1)：38-41

[6] 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

[7] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：165-168

[8] 解冲锋，李星. 基于序列的文本自动分类算法[J]. 软件学报，2002，13(4)：783-788

[9] 韩家炜，孟小峰，王静，等. Web 挖掘研究[J]. 计算机研究与发展，2001，38(4)：405-411

[10] 杜树新，吴铁军. 模式识别中的支持向量机方法[J]. 浙江大学学报：工学版，2003，37(5)：521-527

[11] 罗印升，李人厚，张雷，等. 人工免疫算法在函数优化中的应用[J]. 西安交通大学学报，2003，7(8)：840-843

[12] 王国才，张聪. 一种基于粗糙集的特征加权朴素贝叶斯分类器[J]. 重庆理工大学学报：自然科学版，2010，24(7)：86-90

(上接第 190 页)

结束语　本文介绍了不平衡调节因子和模糊隶属度形成的一种平衡模糊支持向量机处理高不平衡比的数据集的分类问题。模糊隶属度考虑了每个样本对分类超平面的贡献率，从所构建的隶属度的定义可以看出，这种模糊隶属度对一些野点或者异常点的影响具有很好的消除作用。当正类与负类的样本数相差较大时，引入类不平衡调节因子可以根据样本点的分散程度、样本数的多少进行调节，这对不平衡数据集的分类问题具有很好的作用。总之，平衡调节模糊支持向量机较好地处理了不平衡数据集的分类问题，提高了分类精度，但也消耗了大量的时间。

参 考 文 献

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

[2] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer, 1995

[3] Japkowicz N, Stephen S. The Class Imbalanced Problem: A Systematic Study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449

[4] 邹权，郭茂祖，刘扬，等. 类别不平衡的分类方法及在生物信息学中的应用[J]. 计算机研究与发展，2010，47(8)：1407-1414

[5] 吴洪兴，彭宇，彭喜元. 适用于不平衡样本数据处理的支持向量机方法[J]. 电子学报，2006，34(12)：2395-2398

[6] Zeng Zhi-qiang, Gao Ji. Improving SVM Classification with Imbalance Data Set [J]. Neural Information Processing, 2009, 5863: 389-398

[7] 刘万里，刘三阳，王金艳. 不平衡支持向量机的调整方法[J]. 计算机科学，2009，36(3)：148-150

[8] 陶晓燕，姬红兵，马志强. 基于样本分布不平衡的近似支持向量机[J]. 计算机科学，2007，34(5)：174-176

[9] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines[C]// Proc. Int. Joint Conf. Artif. Intell. Stockholm, Sweden, 1999: 55-60

[10] Batuwita R, Palade V. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(2): 558-572

[11] 袁亚湘. 最优化理论与方法[M]. 北京：科学出版社，1997

[12] Huang Kai-zhu, Yang Hai-qin, King I, et al. The Minimum Error Minimax Probability Machine[J]. Journal of Machine Learning Research, 2004, 5(10): 1253-1286

[13] Asuncion A, NewmanD J. UCI repository of machine learning databases [EB/OL]. Http://www. Ics. Uci edu/mlear