一种面向不平衡数据的结构化 SVM 集成算法

袁兴梅, 杨 明

(南京师范大学计算机科学与技术学院, 江苏 南京 210046)
(江苏省信息安全保密技术工程研究中心, 江苏 南京 210046)

[摘要] 不平衡数据在实际应用中广泛存在, 如何处理不平衡数据成为目前一个新的研究热点. 鉴于最大间隔思想在很多分类问题中的优越性, 将最大间隔思想引入到非平衡分类问题中, 使用 SVM 的方法取得了很好的分类性能. 本文在利用类间分布信息的同时, 加上类内结构信息, 使用结构化的 SVM 作为基分类器, 进行分类集成. 实验表明该方法可对不平衡数据进行有效的分类.

[关键词] 不平衡数据, 结构化, 支持向量机, 集成学习
[中图分类号] TP181 [文献标识码] A [文章编号] 1001-4616(2010)04-0123-05

A Kind of StASVM Ensemble Algorithm for Unbalanced Data Sets

Yuan Xingmei, Yang Ming

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China)
(Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210046, China)

Abstract: Imbalanced data sets, arising pervasively in practical application, have attracted more and more attentions. In view of the superiority of maximum margin in many classification problems, we use it for the classification of imbalanced data. Using support vector machine, a good classification performance can be obtained. Based on StASVM, which uses not only between-class information, but also the in-class information, we propose the EStASVM by integrating the obtained subclassifiers induced by StASVM. Experimental results show this ensemble model can better handle the imbalanced problem.

Key words: imbalanced data; structure; SVM; ensemble learning

在模式识别实际应用领域, 经常需要处理类别不平衡问题, 所谓类别不平衡问题是指某类样本数量明显少于其他类样本的情况, 如: 入侵检测, 信用卡交易, 基因编码信息发现等等. 标准的机器学习分类方法在处理不平衡数据分类问题时, 分类判决总会倾向多样本类, 导致少样本类分类精度很低. 因此, 样本数量不平衡导致的分类面偏移成为了分类算法应用于实际问题的主要障碍之一.

国内外学者对不平衡数据分类问题做了大量的研究工作, 提出了许多不同的处理方法: 重采样[1-3]、分类器集成[4-6]、代价敏感学习[7-9]、特征选择[10, 11]、支持向量机[12-14]等. 这些策略可以概括为两类: 一类从训练集入手, 通过改变训练集样本分布, 降低不平衡程度; 另一类是从学习算法入手, 根据算法在解决不平衡问题时的缺陷, 适当修改算法使之适应不平衡分类问题.

支持向量机(SVM)是 Vapnik[15]等人提出的一种分类器, 它主要针对平衡的两类问题, 利用最大间隔的思想有效解决了很多分类问题, 是主流的分类方法之一. 影响 SVM 分类器效果的是总样本中少量的支持向量样本, 因此, 在众多不平衡分类方法中 SVM 似乎更适合处理不平衡数据. 但是, SVM 并不能直接有效地解决此类问题. Wu[16]等人首先洞悉了这样一个事实: 不平衡分类问题可视为介于两类分类与单类分类之间的一种特殊情形, 提出了不平衡支持向量机(ASVM), 其巧妙地利用最大化负类和正类核心间隔的

收稿日期: 2010-07-10
基金项目: 国家自然科学基金(60873176)、江苏省自然科学基金(BK2008430).
通讯联系人: 杨 明, 博士, 教授, 博士生导师, 研究方向: 数据挖掘, 机器学习, 粗集理论与应用. E-mail: yxmnjnu@126.com

思想,同时最大化两类间隔和单类间隔,有效地解决了部分类不平衡问题.通过研究发现,分类器的分类精度同时也与数据类内分布有关,引入数据类内结构信息不仅能提高分类器的分类精度,还能提高分类算法的适用范围.张[17]等人提出了结构化不平衡支持向量机(StASVM),这种算法不仅考虑了类间的分布信息,同时将类内结构信息添加到不平衡支持向量机中,实验表明该算法可以有效提高不平衡数据分类的AUC面积.

而近年来集成学习方法在研究中的有效性越来越明显,当子分类器具有较高的正确率且具有差异性时,通过集成学习可以显著地提高学习系统的泛化性能[18].Tao[19]等人提出了基于 SVM 的 Bagging和随机子空间两种集成方法,这种将 SVM 和集成方法的有效结合可以在图像恢复上取得良好的效果.Liu[20]等人在不平衡数据集上使用欠采样和集成的方法,提出了 EasyEnsemble和 BalanceCascade两种基于欠采样的集成方法.实验表明,这两种方法比先前不平衡问题解决方法在 AUC等方面有明显的提高.可见,集成方法、欠采样、SVM 的有效融合能够取得不错的实验效果.

结构化不平衡支持向量机算法虽然较之前的 SVM 算法有明显的优势,但是这种算法并不稳定.如果数据本身含有较多的噪声或者数据分布杂乱难以聚类时,采样对稳定性的影响较大.基于 StASVM 所存在的问题,本文在其基础上,采用多分类器集成的方法克服单个 SVM 分类器带来的不稳定、泛化能力弱的缺点,可提高不平衡数据的分类性能,增强分类的稳定性.

1 相关知识介绍

1.1 结构化不平衡支持向量机

结构化不平衡支持向量机(StASVM)[17]的出发点是:结合 ASVM 的思想,将结构信息作为正则化项直接嵌入到 ASVM 分类器中,以达到同时考虑类间结构和类内结构的目的.StASVM 算法主要分为两个步骤:首先进行数据聚类,挖掘出数据类内的结构信息,提取出最重要的类内协方差信息,然后将得到的结果信息作为正则化项直接嵌入到 ASVM 的目标函数中.StASVM 的目标函数为:

$$\begin{aligned} &\underset{\omega,\rho,\gamma,\xi}{\operatorname{argmin}} \frac{1}{2}\|\omega\|^2+\frac{\lambda}{2}\omega^{\mathrm{T}}\sum\omega-\rho-\frac{\mu}{\tau}\gamma+\frac{1}{TN}\sum_{i=1}^{N}\xi_i,\\ &s.t.\ \ y_i(\omega^{\mathrm{T}}x_i-\rho)+\frac{1}{2}(y_i-1)\gamma\geqslant-\xi_i,\quad \forall i=1\cdots N,\\ &\xi_i\geqslant 0,\quad \forall i=1\cdots N,\quad \gamma\geqslant 0, \end{aligned} \tag{1}$$

其中 $\sum$ 表示各簇的协方差之和,参数 $\lambda>0$是对最大化间隔和最小化类内紧性的平衡.通过拉格朗日乘子法,(1)的对偶形式可化为:

$$\begin{aligned} &\underset{\partial}{\operatorname{argmax}}\ \partial^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{I}+\lambda\sum\right)^{-1}\boldsymbol{X}\boldsymbol{Y}\partial\\ &s.t.\ \ \partial^{\mathrm{T}}1\geqslant 2\frac{\mu}{\tau}+1,\\ &\partial^{\mathrm{T}}y=1,\quad 0\leqslant\partial\leqslant\frac{1}{TN}, \end{aligned} \tag{2}$$

其中 $\boldsymbol{X}=[x_1,\cdots,x_N]$, $\partial=[\partial_1,\cdots,\partial_N]^{\mathrm{T}}$, $\boldsymbol{Y}=\mathrm{diag}(y_1,\cdots,y_N)$,而 1表示 N 维的全 1列向量.将(2)的第一个约束条件变为 $\partial^{\mathrm{T}}1=2\mu/\tau+1$,此时即可通过 SMO求解得到法向量.

这种算法不仅同时最大化类间间隔和单类间隔,更重要的是通过引入类内结构信息,实现了最大化类内紧性,融入了更多的先验知识,由此可保证 StASVM 有更好的分类性能.当线性 StASVM 不能有效地分开两类数据时,同样也可以通过核技巧解决非线性问题.

1.2 集成学习

传统的学习方法是在一个由各种可能的函数构成的空间中寻找一个最接近实际分类函数 f 的分类器 h.单个分类器模型主要有决策树,人工神经网络,支持向量机等等.但是单个学习器的泛化能力不足:训练集并不能提供足够的信息来选择一个最好的分类器;学习算法的搜索过程并不总是最优的;被搜索的假设空间可能不包含真正的目标函数[21].

集成学习在对新的实例进行分类的时候,把若干单个分类器结合起来,通过对多个子分类器的分类结

果进行某种组合来决定最终的分类, 以取得比单个分类器更好的性能. 它可以概括为两大步: 基分类器的获取和基分类器分类结果的整合. 要获得好的集成效果, 提高基分类器之间的差异性是很重要的一个方面, 人们从划分训练集或特征集的角度, 总结出了几大方法: Bagging, Random Subspace, Adaboost. 集成学习的方法如果使用得当, 能够有效地避免单个学习器自身所固有的缺点和局限. 但并不是所有的集成学习都有效, 其有效的条件是每个单一的学习器错误率都应当低于 0.5, 而且单个分类器之间要具有一定的差异性, 否则集成的结果反而会提高分类错误率[21,22].

2 面向不平衡数据的 *ESt*A*SVM* 算法

本文采用基于欠采样的方法, 运用 StA SVM 分类模型进行集成分类器设计. 设计的主要思路: (1) 基于聚类欠采样, 得到各子分类器的训练数据集; (2) 由各训练数据集诱导出子分类器; (3) 采用加权策略得到集成分类器 ESt A SVM. 具体算法描述如下:

输入: 训练数据集 $D = \{(x_j, y_j) \mid 1 \leqslant j \leqslant N\}$, $x_j \in X$, $y_j \in C = \{-1, +1\}$;
基分类器个数 basenum;
聚类后每个簇的采样率 rate_ per_ clu;
变量描述: trainpositive: 正类训练样本;
trainnegative: 负类训练样本;
testdata: 测试样本;
cluster_ negative: 对大类样本进行聚类后带簇标号的样本;
clusternum: 对大类样本聚类后簇的个数;
sam_ negative: 聚类后对大类的各簇欠采样后得到的样本;
输出: 集成分类器 H 对测试数据的平均 AUC面积.

```
For n = 1: 10 do
    (trainpositive, trainnegative, testdata) = Randperm(1, 9);
    /* 抽取数据集中正、负类样本符合 1: 9的比率作为训练样本 */
(cluster_ negative, clusternum) = clusterdata(trainnegative);
    /* 对大类的训练样本进行 ward层次聚类 */
    for i = 1: basenum do
      for j = 1: clusternum do sam_ negative(j) = undersampling(cluster_ negative,
        rate_ per_ clu);
        /* 聚类后对各簇进行欠采样 */
      end
      (fx(i), AUC(i)) = StASVM(sam_ negative, trainpositive, test);
        /* 用欠采样后的样本和小类样本训练结构化的非平衡支持向量机 */
    end
      rate = account(AUC);
      /* 根据训练到的每个基分类器的 AUC计算每个基分类器在决策中的权重 */
      (resultfx(n), resultAUC(n)) = Ensemble(rate, fx);
      /* 对本轮训练得到的基分类器集成, 得到集成后测试样本的 AUC 面积 */
End
averageAUC = mean(resultAUC) /* 计算该数据集轮数为 n的平均 AUC* /
```

在 ESt A SVM 中, 使用了与 SLMM[23] 和 SSVM[24] 相同的 W ard层次聚类算法, 该聚类算法每次将两个距离最小的聚类合并成一个更大的聚类, 随着聚类个数的减小, 合并距离随之增大, W ard聚类方法选择在合并距离曲线的拐点处结束聚类过程. 这种聚类方法不需事先确定聚类簇的个数, 而是通过合并距离曲线的拐点自动停止聚类合并过程. 由于本文处理的是不平衡数据, 正类规模较小, 不进行聚类; 我们将规模较大的负类通过层次聚类聚成了 c 个簇, 分别记为 $N_1, \cdots, N_c$.

进一步, 在 ESt A SVM 中, 采用基于聚类的欠采样. SVM 在样本集分布均衡的情况下, 能够取得较好的分类效果, 但在样本集不平衡情况下, 对少类样本分类效果较差. 通过已有的大量研究[1,2,25] 发现, 使用欠采样技术设计 SVM 分类器有这样几个优点: ①由于 SVM 中涉及复杂的核矩阵求解, 计算复杂度较高, 而

欠采样方法能够减小样本集容量,因此能够大大缩短计算时间;②具有较为明确的物理含义,便于理解和编程实现.因此在 SVM分类器设计中,非常适合使用欠采样技术[2].本文采用基于聚类的欠采样技术的目的在于:在尽量缩小两类样本数量差异的前提下,最大化地保留原样本内部的分布信息,在一定程度上缓解由于样本集失衡造成的分类器整体效果变差的问题.因大多数文献[2,19,20]使用的欠采样多是基于随机采样,没有考虑多类样本集内部真实分布特性,使得采样后的数据集容易造成分类信息严重丢失.而在聚类基础上的欠采样,可以去除各簇中相对冗余的样本,又能有效避免盲目抽样带来的信息丢失.此外,这种方法的欠采样还可以消除部分噪声样本,对于单个样本形成的簇,欠采样率小于 1则可以去除该样本.当欠采样率为 1时,EStASVM算法的基分类器就退化为 StASVM.

基于聚类的欠采样完成之后,可得到抽样后每个簇的协方差矩阵 $\sum_{N_1}$, …, $\sum_{N_c}$, $\sum_{N_p}$.然后将这些结构信息融入最大间隔分类器设计中:利用总体协方差矩阵构造出表示最大化类内紧性的正则化项,并将它直接嵌入 ASVM的目标函数中用来训练得到最终的基分类器 StASVM(*i*).在对每一个基分类器训练时,由于本文处理的是不平衡数据,AUC是度量模型判别能力的重要指标.在 StASVM参数选取时,采用网格搜索的方法,以 AUC最大化为训练目标,为各基分类器选取合适的参数.

3　实验结果

为了验证本文算法的有效性,我们在 UCI数据集上进行实验.另外,由于需要极度不平衡的样本,但 UCI数据集不满足实验需求,所以需要对数据集进行采样或者进行类合并操作.本实验中正类样本数和负类样本数之比为 1∶9,训练样本和测试样本之比约为 1∶1.表 1已列出训练过程中正负类样本的规模.

表 1　实验中使用的数据集和样本规模

Table 1　Data sets and sample size used in experiment

	正类样本总数/正类训练样本数	负类样本总数/负类训练样本数
WDBC	36/18	324/162
bupa	20/10	180/90
heart	16/10	144/90
wine	13/8	117/72
glass	16/10	144/90
ionosphere	25/13	225/117
sonar	12/6	108/54

实验中选择高斯核函数,参数采用网格搜索方式,每组数据集重复 10轮,实验结果取 10轮的平均值.

表 2是这 7个数据集在 ASVM, StASVM, EStASVM这三大方法中的 AUC结果,实验结果是 10轮实验的均值.表 3是分别使用 StASVM和 EStASVM这两种算法 10轮实验 AUC结果的标准差,反映的是这两种算法的实验稳定性.由表 2和表 3,可以发现:

①加入了类内结构信息,StASVM在 AUC方面比 ASVM有较大的提高,继而在此基础上进行集成学习,EStASVM的效果又比单纯的加进结构信息有了进一步改进,尤其是 Bupa和 sonar数据集.

②由表 2可见,在 ionosphere数据集上使用集成方法,效果反而不如 StASVM.通过多次实验发现这可能因为在该数据集上聚类后,不少样本没有能和其他样本合并,作为单独的一个簇作为最终的聚类结果.这样在此基础上进行欠采样,只要采样率不为 1,则该簇就没有样本参加 SVM模型的训练,也就是说该簇的信息完全丢失.如果该样本原本是支持向量,那么使用 EStASVM算法反而丢失了重要信息.

③由表 3可见,StASVM算法不够稳定性,这是因为在对数据集进行采样时,如果样本集含有较多的噪声和数据分布杂乱而不易分类时,采样对稳定性的影响较大.但通过集成的添加,除 ionosphere之外,稳定性均有了不同程度的提高.

表 2　ASVM、StASVM、EStASVM在 UCI数据集上的 AUC实验结果

Table 2　Experiment results with algorithm ASVM, StASVM and EStASVM

	ASVM	StASVM	ESTASVM
WDBC	0.9650	0.9711	0.9786
Bupa	0.6684	0.7175	0.7513
Heart	0.7225	0.7297	0.7419
Wine	0.8911	0.9256	0.9533
Glass	0.8613	0.8933	0.9041
ionosphere	0.9574	0.9718	0.9587
sonar	0.7351	0.7544	0.7932

表 3　StASVM和 EStASVM实验稳定性比较

Table 3　Comparison of experiment stability between StASVM and EStASVM

	StASVM	ESTASVM
WDBC	0.00098	0.00053
Bupa	0.07556	0.01128
Heart	0.03147	0.00112
Wine	0.00711	0.00519
Glass	0.00754	0.00573
ionosphere	0.00064	0.00081
sonar	0.06842	0.00842

为了进一步验证聚类后结构信息的加入和在此基础上采样对实验效果的影响, 本文以 bupa数据集为例, 给出了 3次随机采样后负类样本的聚类情况, 以及在聚类基础上不同采样率的 AUC结果, 如图 1所示. 由图 1可见, bupa数据集在采样率为 0.3时 AUC值达到最大; 而由聚类簇的个数结果来看, bupa在聚成 5簇时, AUC面积值最大. 所以 AUC结果与聚类结果和抽样比例的关系密切.

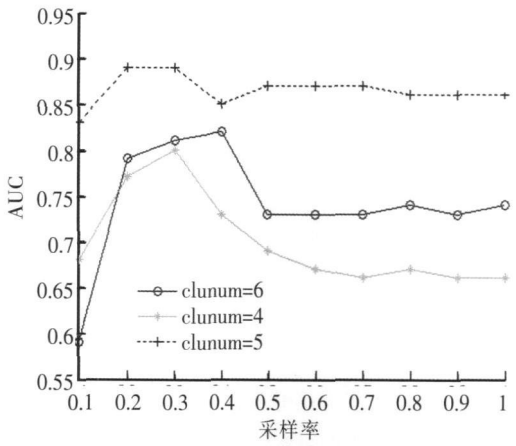

图 1 bupa 数据集聚类结果及采样率比例与 AUC 的关系

Fig.1 The relationship between sampling rates and AUC of bupa dataset under different clustering results

4 结语

对于不平衡问题的处理, 本文提出了一种基于结构化支持向量机的集成学习算法 EStASVM. 实验效果表明, 该算法在处理不平衡数据时比使用单分类器时分类效果和稳定性上都有了显著提高, 但实验也表明, 该算法的性能依赖于 Ward聚类算法的性能. 在未来的工作中我们需要结合多种方法设计适应性更广的分类器, 对于多类不平衡的情况以及如何优化 ward算法也是未来的研究目标之一.

[参考文献]

[1] Batisa Geapa, Pratirc, Monard M C. A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations Newletter, 2004, 6(1): 20-29.

[2] Barandelar, Valdovinos R M, Sanchez J S, et al. The imbalanced training sample problem: under or over sampling[C] // Proc of International Workshops on Str-uctural, Syntactic, and Statistical Pattern Recognition. Lisbon, 2004.

[3] Chawlanv, Hallh, Bowyer K W, et al. SMOTE: synthetic minority oversampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(6): 321-357.

[4] Fan Wei, Stolfo S J, Zhang Junxin, et al. AdaCost: misclassification cost-sensitive boosting[C] // Proceedings of the 16th International Conference on Machine Learning. 1999.

[5] Joshi M, Kumar V, Agarwal R. Evaluating boosting algorithms to classify rare classes: comparison and improvements[C] // Proceedings of the First IEEE International Conference on Data Mining, 2001.

[6] Liu Yang, An Aijun, Huang Xiangji. Boosting prediction accuracy on imbalanced datasets with SVM ensembles[C] // Proceedings of the 10th Pacific-Asia Conference on Adavances in Knowledge Discovery and Data Mining. Berlin, 2006.

[7] Pazzanim, Merz C, Murphy P, et al. Reducing misclassification costs[C] // Proceedings of the 11th International Conference on Machine Learning. San Francisco, 1994.

[8] Yanmin Sun, Mohamed S Kamel, Andrew K C Wong, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(2): 3 358-3 378.

[9] Chawla N V. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure[C] // Proceedings of International Conference on Machine Learning. Washington DC, 2003.

[10] Cardie C, Howen. Improving minority class predicting using case-specific feature weighted[C] // Proceedings of the 14th International Conference on Machine Learning. San Francisco, 1997.

[11] Zheng Z H, Harir S R. Optimally combining positive and negative features for text categorization[C] // Proceedings of International Conference on Machine Learning. Washington DC, 2003.

[12] Japkowicz N, Stephen S. The class imbalance problem: a systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 203-231.

[13] Brefeld U, Scheffer T. AUC maximizing support vector learning[C] // Proceedings of International Conference on Machine Learning Workshop on ROC Analysis in Machine Learning. Bonn, 2005.

[14] Amari S, Wu S. Improving support vector machine classifiers by modifying kernel functions[J]. Neural Networks, 1999, 12(6): 783-789.

[15] Vapnik V. Statistical Learning Theory[M]. New York: John Wiley and Sons, 1998.

(下转第 133页)

宽度, 用检测到的结果构造数学形态学线性结构元素, 用数学形态学运算实现道路的提取. 结构元素方向和尺度两个参数根据图像本身自适应得到. 该方法使结构元素方向随遥感影像中主体道路方向变化, 保证了所有主要方向上的道路都能提取; 结构元素尺度随遥感影像主体道路宽度变化, 保证了合适尺度的道路能被准确提取. 实验结果表明, 用 Hough变换得到的道路特征来构造结构元素能够使数学形态学更好地提取高分辨率图像中的城区道路. 对于城区道路中存在的立交桥等曲率比较大的道路, 本文提出的方法还不能很好地提取, 这将是下一步的研究工作.

[参考文献]

[1] Ma Hongbin, Zhao Yahong, He Qun. Road extraction from high resolution remote sensing image based on mathematics morphology and seed growth[C]// International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2008: 523-526.

[2] Letitia S, Elwin Chandramonie. Segmentation of urban road network from satellite images using fuzzy mathematical morphology[J]. ICGST-GVIP, 2008, 8(4): 27-32.

[3] 潘建平, 邬明权. 基于数学形态学的道路提取[J]. 计算机工程与应用, 2008, 44(11): 232-233.

[4] 安如, 冯学智, 王慧麟. 基于数学形态学的道路遥感影像特征提取及网络分析[J]. 中国图像图形学报, 2003, 8(7): 798-804.

[5] 刘生, 王潇宇. 基于数学形态学的高空间分辨率遥感影像几何特征提取[J]. 地球信息科学, 2008, 10(2): 251-256.

[6] 吕健刚, 韦春桃. 基于 Hough变换的高分辨率遥感影像城市直线道路提取[J]. 遥感应用, 2009(3): 15-18.

[7] Bong D B L, Lai K C, Joseph A. Automatic road network recognition and extraction for urban planning[J]. World Academy of Science, Engineering and Technology, 2009, 41(53): 209-210.

[8] 徐春燕, 冯学智, 赵书河, 等. 基于数学形态学的 IKONOS多光谱图像分割方法研究[J]. 遥感学报, 2008, 12(6): 980-986.

[责任编辑: 孙德泉]

(上接第 127页)

[16] Wu S H, Lin K P, Chen C M. Asymmetric support vector machines: low false positive learning under the user tolerance [C]// Proceeding of the 14th ACM SIGK-DD International Conference on Knowledge Discovery and Data Mining. New York, 2008.

[17] 张青青, 陈松灿. 非平衡类的异常检测研究[D]. 南京: 南京航空航天大学信息科学与技术学院, 2010.

[18] Zhou Zhihua, Li Nan. Multi-information ensemble diversity[C]// Proceedings of the 9th International Workshop on Multiple Classifier Systems. Cairo Egypt, 2010.

[19] Tao Dacheng, Tang Xiaoou, Li Xuelong, et al. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1 088-1 099.

[20] Liu Xuying, Wu Jianxin, Zhou Zhihua. Exploratory undersampling for class imbalance learning[J]. IEEE Transactions on Systems, Man and Cybernetics-part B: Cybernetics, 2009, 39(2): 539-550.

[21] Huang Faliang, Xie Guoqing, Xiao Ruliang. Research on ensemble learning[C]// Proceeding of the International Conference on Artificial Intelligence and Computational Intelligence. Shanghai, 2009.

[22] Yu Lean, Wang Shouyang, Kin Keung Lai. Investigation of diversity strategies in SVM ensemble learning[C]// Proceedings of the 4th International Conference on Natural Computation. Jinan, 2008: 39-42.

[23] Yeung D S, Wang D, Ng W W Y, et al. Structured large margin machines: sensitive to data distributions[J]. Machine Learning, 2007, 68(2): 171-200.

[24] Xue H, Chen S, Yang Q. Structural support vector machines[C]// Proceedings of the 15th International Symposium on Neural Networks. Beijing, 2008.

[25] He Haibo, Edwardo A Garcia. Learning form imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1 263-1 284.

[责任编辑: 孙德泉]