不平衡数据分类的研究现状 *

林智勇[1a,2], 郝志峰[1a], 杨晓伟[1b]
(1. 华南理工大学 a. 计算机科学与工程学院; b. 数学科学学院, 广州 510641; 2. 广东技术师范学院 计算机科学系, 广州 510665)

摘 要: 不平衡数据在实际应用中广泛存在,它们已对机器学习领域构成了一个挑战,如何有效处理不平衡数据也成为目前的一个新的研究热点。综述了这一新领域的研究现状,包括该领域最新研究内容、方法及成果。

关键词: 不平衡数据; 机器学习; 模式分类
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2008)02-0332-05

Current state of research on imbalanced data sets classification learning

LIN Zhi-yong[1a,2], HAO Zhi-feng[1a], YANG Xiao-wei[1b]
(1*a*. *School of Computer Science & Engineering*, *b*. *School of Mathematics*, *South China University of Technology*, *Guangzhou* 510641, *China*; 2. *Dept. of Computer Science*, *Guangzhou Polytechnic Normal University*, *Guangzhou* 510665, *China*)

Abstract: IDS(imbalanced data sets), arising pervasively in practical applications, have caused a huge challenge to the machine learning community and consequently attracted more and more attentions. The paper summarized the state of this relatively new research field, including its issues, methods and results.

Key words: imbalanced data sets(IDS); machine learning; pattern classification

不平衡数据分类考虑的是各类样本数目不平衡情况下的分类学习问题。以二分类为例,若其中有一类(正类、多数类)的学习样本比另一类(负类、少数类)的学习样本多得多,那么就称这样的分类问题为 IDS 分类问题。IDS 在实际应用中经常碰到,如欺诈识别、入侵检测、医疗诊断以及文本分类等都是典型的 IDS 问题。传统的分类方法主要考虑的是各类学习样本数量大致均衡的情形,其评价标准主要是基于精度的。这使得现有的分类方法往往不能有效地处理 IDS,尤其是数据的不平衡严重时(正/负类学习样本数量比可高达100∶1、1 000∶1甚至10 000∶1)更是如此。

1 IDS 分类学习的应用

随着应用的广泛深入,人们发现 IDS 分类问题并非少见,许多实际问题的数据是不平衡的。早在 1998 年, Kubat 等人[1]就考虑了一个实际的 IDS 分类问题,他们根据获得的卫星图像,通过分类的方法对石油喷井进行计算机自动监测。其中数据不平衡的比例约为22∶1。此后, Phua 等人[2]和 Pérez 等人[3]相继考虑了欺诈识别中的 IDS 分类问题。文献[3]的数据集中具有108 000个样本,只有约 7.4% 是欺诈样本。Castillo 等人[4]以及 Zheng 等人[5]研究了文本分类问题;Cohen 等人[6]考虑了医院传染病监测问题;Chen 等人[7]讨论了 IDS 分类学习在药物治疗检测方面的应用;Yoon 等人[8]介绍了在生物信息学方面的应用;Radivojac 等人[9]则将 IDS 分类学习应用于无线传感器网络的入侵检测方面。在诸多实际应用中,研究者们均指出了数据不平衡对分类学习带来的困难和挑战,其中最主要的方面就是分类器性能大大降低。

2 IDS 问题实质探讨

不平衡问题的实质是什么? 各类学习样本数量的不均衡是否一定会降低传统分类方法的性能? 进一步地,影响分类器性能的因素有哪些,这些因素对各种不同的分类方法的影响是否相同? 这些都是 IDS 给人们带来的新的思考。Japkowicz 等人[10]通过实验的方法对 IDS 问题进行了较为系统的研究,她考虑了概念复杂度、训练样本规模和类间不平衡程度三个因素对分类器性能的影响。实验表明,除了类间不平衡程度(即类间学习样本数量比例)这个因素外,另外两个因素也会对分类器性能产生影响。当概念复杂度较低时,类间不平衡程度并不会对分类器性能产生太大的影响;此外,提高训练样本规模也可缓解类间不平衡对分类器性能的不良影响。文献[10]还比较了数据不平衡对不同分类方法的影响。其中包括基于决策树的 C4.5、BP 神经网络以及支持向量机(SVM)等。实验结果表明,相对而言,SVM 对数据不平衡带来的影响较不敏感。在此基础上, Jo 和 Japkowicz[11]进一步比较研究了类间(between-class)不平衡和小析取项(small disjuncts)对分类学习的影响。小析取项即类内(within-class)不平衡,从概念学习的角度来说,它反映了同一类的若干子概念之间学习样本分布的不平衡性。小析取项就是那些所涵盖的学习样本数量偏少的子概念,它们是容易被错误学习进而影响分类器整体性能的一个重要因素。类间和类内不平衡是 IDS 的两个不同侧面,它们可能会同时出现,均会影响分类器的性能。Prati 等人[12]挑选了

收稿日期: 2007-01-16; **修回日期**: 2007-03-26 **基金项目**: 国家自然科学基金资助项目(60433020,10471045)

作者简介: 林智勇(1977-),男,广东梅州人,博士研究生,主要研究方向为机器学习、计算智能等(zy_lin@21cn.com);郝志峰(1968-),男,江苏苏州人,教授,博导,博士,主要研究方向为代数学及其应用、计算智能、机器学习等;杨晓伟(1969-),男,河南平顶山人,副教授,硕导,博士,主要研究方向为机器学习等.

UCI[13]的十个数据集作为实验数据,对这两种不平衡进行了实验比较研究,此外,Prati 等人[14]还对类不平衡和类重叠进行了比较研究。他们指出分类器性能的下降不能只归咎于类不平衡的存在。在一些类严重不平衡的分类学习中,分类器仍然具有良好的性能。这是因为类重叠并不严重,也就是说,类重叠也是影响分类器性能的一个重要因素。

3　分类器的合理评价

在传统的分类学习方法中,训练精度是主要的评价指标。然而对于 IDS 问题来说,用精度来评价分类器的性能却并不合理。比如在二分类中,假设正类的样本占了 99%。若本文的分类方法就是将所有的样本都归为正类,那么这个简单的分类器就可以获得高达 99% 的训练精度。但是这样的分类器是没有实用价值的。对于一个具体的分类器,考虑如表 1 所示的混淆矩阵。其中:Pos 表示正类样本;Neg 表示负类样本;N = Pos + Neg 为全体学习样本;TP(true positive)和 TN(true negative)分别表示被正确分类的正类和负类样本;FP(false positive)和 FN(false negative)则分别表示被错分的正类和负类样本。根据这个矩阵,可以定义如下的量[15]:

TP rate = TP/Pos = TP/(TP + FN)

FP rate = FP/Neg = FP/(TN + FP)

精度 accuracy = (TP + TN)/N

查准率 precision = TP/(TP + FP)

查全率 recall = TP rate

显然,查全率和查准率都是越大越好,而 FP rate 则越小越好。对分类器的合理评价应该综合考虑这些指标。接收者操作特性(receiver operating characteristic,ROC)考虑的是 TP rate 和 FP rate。在 ROC 空间中,以 FP rate 为横轴、TP rate 为纵轴对分类器进行定位,如图 1 所示[15]。

表 1　混淆矩阵

系 数	predicted positive	predicted negative	
positive examples	TP	FN	Pos
negative examples	FP	TN	Neg
	PPos	PNeg	N

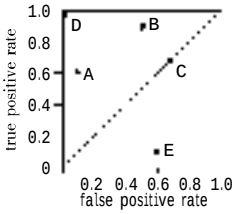

图 1　ROC 示意图

ROC 空间中的每一个点对应于一个具体的分类器。显然,越是位于左上角的分类器其性能越好。在图 1 中,D 所对应的分类器性能最理想。对于某些分类方法(如神经网络),通过调整阈值可以在 ROC 空间中得到一族点,连接这些点便形成所谓的 ROC 曲线。若一个分类方法的 ROC 曲线总是在另一个方法的 ROC 曲线的上方,那么前者要比后者好。然而,许多时候两条 ROC 曲线会有交叉,因此采用曲线下方图面积 AUC(area under ROC curve)作为评价标准,AUC 越大越好。事实上,AUC 具有统计意义[16]。假设分类决策是根据函数 $f(x)$ 值进行的,且正确分类情形下正类样本对应的函数值大于负类样本的函数值,那么 $\mathrm{AUC}(f) = P(f(x^+) > f(x^-))$,即对于随机抽取的一个正类样本 x^+ 和负类样本 x^-,f 赋予 x^+ 比 x^- 更大的函数值的概率。

Drummond 等人[17]将误分代价考虑进去,提出所谓的 cost 曲线。定义如下的量:

$$E[\mathrm{cost}] = \mathrm{FN} \times P(+) \times C(-|+) + \mathrm{FP} \times P(-) \times C(+|-)$$

$$\max E[\mathrm{cost}] = P(+) \times C(-|+) + P(-) \times C(+|-)$$

$$\mathrm{Norm}(E[\mathrm{cost}]) = E[\mathrm{cost}]/\max E[\mathrm{cost}] = \mathrm{FN} \times PC(+) + FP \times PC(-)$$

$$PC(+) = P(+) \times C(-|+)/[P(+) \times C(-|+) + P(-) \times C(+|-)]$$

$$PC(-) = P(-) \times C(+|-)/[P(+) \times C(-|+) + P(-) \times C(+|-)]$$

其中:$P(+)$ 和 $P(-)$ 表示两类的先验概率;$C(-|+)$ 和 $C(+|-)$ 分别表示误分代价;$E[\mathrm{cost}]$ 的意义是分类器的平均误分代价;$\mathrm{Norm}(E[\mathrm{cost}])$ 是归一化之后的平均误分代价;$PC(+)$ 和 $PC(-)$ 则可以理解成考虑了误分代价的两类先验概率。在 cost 空间中,横轴表示 $PC(+)$,纵轴则表示 error rate(即 $\mathrm{Norm}(E[\mathrm{cost}])$)。Drummond 等人对 cost 空间和 ROC 空间进行了比较研究,指出两者之间存在某种对偶关系。他们认为在 cost 空间中可以更方便地进行分类器性能的评价和比较。

ROC 以及 cost 曲线将分类器性能可视化,其优点是直观明了,但不够方便。人们往往需要某些量化指标(如 AUC),这样使用起来更方便,而且也易于作为分类器优化的标准。针对 IDS 分类问题,人们主要考虑了以下指标[8,19]:

a)两类训练精度的几何平均 $(acc^+ \times acc^-)^{1/2}$。其中:$acc^+$ = TP/(TP + FN);acc^- = TN/(TN + FP)。

b)查准率和查全率的几何平均 $(\mathrm{precision} \times \mathrm{recall})^{1/2}$。

c)F-得分:$F_\beta = (\beta^2 + 1)\,\mathrm{precision} \times \mathrm{recall}/(\beta^2 \mathrm{precision} + \mathrm{recall})$。其中:$\beta \geqslant 0$ 是参数,常选择为 1。显然,F_0 = precision,而 $F_{+\infty}$ = recall。当 $0 < \beta < +\infty$ 时,F_β 在 precision 与 recall 之间进行了某种折中。

4　数据层面的处理方法

数据层面的处理是对数据进行重抽样,包括过抽样和欠抽样两种。其主要思想是通过合理地增加或者减少一些样本去平衡化数据,进而降低数据不平衡对分类器带来的不良影响。最简单的重抽样方法就是随机增加(复制)或删除部分样本,但其效果通常不理想,人们考虑得更多的是启发式的做法。

一般说来,欠抽样主要是去掉噪声和冗余数据,而且主要是针对多数类的样本进行。其中:常用的技术包括 Tomek link、一致子集(consistent subset)、编辑技术(常用的是 Wilson's editing)以及单边选择(one-sided selection)等[19,20]。这些技术主要是启发式地利用(加权)欧氏距离以及 K-近邻规则去识别可以合理剔除的样本。Barandela 等人[21]以及 Batista 等人[19]都对上述的多种欠抽样方法进行了详细的实验比较研究。Dehmeshki 等人[22]则提出了基于规则的数据过滤技术,其本质上也是欠抽样方法。他们通过构造规则去发现安全区域;然后将安全区域内的样本点剔除掉(针对多数类进行)。他们认为其中的道理在于安全区域内的样本对分类器的构建并无太大作用,因此可以剔除。

与欠抽样相反,过抽样技术主要是设法增加少数类的学习样本。其中的代表是由 Chawla 等人[23]提出的 SMOTE 技术。SMOTE 技术的主要想法在于通过插值生成新的人造样本,而不是简单地复制样本。Han 等人[24]在此基础上进行改进,提出了 Borderline-SMOTE 技术。其主要想法是在适当的区域内进行插值,以保证新增加的样本是有价值的。

在实际应用中,为了获得好的效果,经常将不同的欠抽样和过抽样技术混合使用。

5 算法层面的处理方法

针对 IDS 问题改进原有算法或者设计更有效的新算法是 IDS 分类学习研究中最主要的组成部分。根据笔者所掌握的文献资料来看,目前主要集中在如下四个不同的途径:代价敏感(cost-sensitive)学习、支持向量机方法、单类(one-class)学习、组合(combining)方法等。

5.1 Cost-sensitive 学习

Cost-sensitive 学习中考虑了误分代价[25~27],给定一个代价矩阵 C。其中,$C(i,j)$ 表示将类标号为 j 的样本误分为类 i 的代价(通常假定 $C(i,i)=0$)。那么 cost-sensitive 学习最小化的条件风险[25]如下:$R(i|x)=\sum_j P(j|x)C(i,j)$。显然,这里的 $R(i|x)$ 表示将 x 分为类 i 的平均代价。进而对样本 x 而言,应该选择 $k=\arg\min_i R(i|x)$ 作为其类别号。在处理 IDS 问题时,若类 j 是少数类,则通常选择 $C(i,j)>C(j,i)$。其中的合理之处在于:一方面在实际问题中,误分代价通常是不相等的,如在医疗诊断中,将一个患者误判为正常人比将一个正常人误判为患者的代价要大;另一方面,考虑误分代价将是分类边界适当向多数类偏移,从而能提高少数类的分类精度,这对于稀少类的情形是有用的[28]。

5.2 支持向量机方法

根植于统计学习理论之上的 SVM 分类方法有着良好的推广性能,目前已得到广泛的应用。实验研究表明,相对而言,SVM 分类器对数据的不平衡性更不敏感[10]。因此,人们考虑对 SVM 进行适当的改进以更好地处理不平衡数据。一种简单的方法是将分类边界朝多数类进行适当的偏移,以使更多的少数类样本不会被误判:$\langle w,\Phi(x)\rangle+b=0\rightarrow\langle w,\Phi(x)\rangle+b+\Delta=0$。

另一种做法是在 C-SVM 中对不同的类采取相同的惩罚因子[29,30]:

$$\min f=\|w\|^2/2+C^+\sum_{y_i=+1}\xi_i+C^-\sum_{y_j=-1}\xi_j$$

$$\text{s.t. } y_i(w\times\Phi(x_i)+b)\geqslant 1-\xi_i;\ \xi_i\geqslant 0$$

若正类是多数类,则通常选择 $C^+<C^-$。然而,Wu 等人[31]认为上述两种改进的作用不大,他们在 Amari 等人[32]工作的基础上提出核边界校准算法。其主要思想是通过引入一个伪保形(quasi-conformal)函数 $D(x)$ 对原始核函数 $k(x,x')$ 进行变换:

$$k(x,x')\rightarrow\tilde{k}(x,x')=D(x)D(x')k(x,x')$$

实际上,比这更一般的问题是如何选择有效的核函数。

Chen 等人[33]考虑对支持向量进行裁减,通过适当牺牲多数类的分类精度以提高少数类的精度。Brefeld 等人[16]则直接以 AUC 极大化为目标,提出了新的 SVM 形式的分类方法。类似地,Callut 等人[34]以 F-得分 F_β 为准则提出了 F_βSVM。值得指出的是,在文献[16,34]所提出的方法中,虽然其最后优化的问题形式与 SVM 相似,但其中的意义却有所不同。它们都是在优化适用于 IDS 的某个评价指标(如 AUC)基础上得到的,因此在处理 IDS 分类问题上更具有直观意义。

5.3 单类学习

当类间数据严重不平衡时,分类器通常都会倾向于将几乎所有的数据判为多数类。为了解决这个问题,人们考虑采用不是基于区别的分类方法,而是基于识别的方法进行学习,进而提出了单类学习。单类学习方法的主要思想在于只利用感兴趣的目标类的学习样本进行学习。对于新的样本,通过比较该样本与目标类的相似程度而识别该样本是否归属于目标类。在单类学习中,目前研究得比较多的还是基于 SVM 的方法。Schölkopf 等人[35]首先将 SVM 用于密度估计。考虑了如下的单类 SVM:

$$\min f=\|w\|^2/2+1/(vl)\sum_i\xi_i-\rho$$

$$\text{s.t. } w\times\Phi(x_i)\geqslant\rho-\xi_i;\ \xi_i\geqslant 0$$

这里的 x_i 均是来自同一个类别的样本。此后,人们便将单类 SVM 用于单类学习进而解决一些 IDS 分类问题,尤其是当少数类的学习样本非常少的情形下[36~38]。

5.4 组合方法

组合方法的主要思想在于将多个分类器组合成一个分类器,以提高分类性能。其中,提升是被广泛使用的技术。通过提升,多个弱分类器可以组合成一个强分类器。AdaBoost 是采用提升技术算法的代表[39]。在该算法中,最终得到的分类器是多个弱分类器的线性组合形式:

$$H(x)=\text{sign}(\sum a_tH_t(x))$$

算法首先给出各学习样本的初始权重(分布)$D_t(i)$($\sum_i D_t(i)=1$);然后根据这些带权样本进行训练得到相应的弱分类器 $H_t(x)$;接着利用所获得的 $H_t(x)$ 计算相应的组合系数:

$$a_t=\ln[(1+r_t)/(1-r_t)];\ r_t=\sum_i D_t(i)y_iH_t(x_i)$$

其中:$y_i=\pm1$ 是 x_i 的类标号。随后,对样本权重作如下调整:

$$D_{t+1}(i)\leftarrow D_t(i)\exp(-a_ty_iH_t(x_i))/Z_t$$

这里的 Z_t 是归一化因子,使得 $\sum_i D_{t+1}(i)=1$。如此不断地迭代,直到满足终止条件为止。大量的实验证实,通过 AdaBoost 提升所得的弱分类器组合具有良好的分类性能。最近的研究表明,其中的道理在于 AdaBoost 实际上也是一种间隔极大化的算法。它与 SVM 具有相似之处[40]。

鉴于提升技术的简单有效,不少学者将它用于处理 IDS 分类问题。Fan 等人[27]首先将误分代价嵌入 AdaBoost 中得到 AdaCost 算法。其关键之处在于采用如下的权重调整策略:$D_{t+1}(i)\leftarrow[D_t(i)\exp(-a_ty_iH_t(x_i)\beta(i))/Z_t]$。这里,$\beta(i)$ 是与误分代价有关的调整因子。与 Fan 等人的做法类似,Leskovec 和 Shawe-Taylor 也在权重调整中引入因子以反映类不平衡性,提出了 AdaUBoost 算法。进一步地,他们采用线性规划进行提升,获得另一种被称为 LPUBoost 的算法用于处理 IDS[41,42]。Joshi 等人[43]则考虑对正类和负类预测采取不同的组合因子,其最终的分类器形式为 $H(x)=\text{sign}(\sum_{t:H_t(x)>0}a_t^pH_t(x)+\sum_{t:H_t(x)<0}a_t^nH_t(x))$。这样做可以适当兼顾少数类的分类精度,尤其适用于稀有类的情形。他们将所得的算法称为 RareBoost 算法。

除了在形式上改进提升算法之外,人们还将提升与抽样技术相结合以处理 IDS。Chawla 等人[44]将过抽样技术 SMOTE 与提升技术相结合,提出了 SMOTEBoost 算法。类似地,Guo 等人[45]提出了 DataBoost 算法。Liu 等人[46]和 Kang 等人[47]则考虑通过重抽样形成多个训练样本子集,然后利用 SVM 方法进行学习得到多个分类器,然后将这些分类器组合成最终的分类器。

5.5 其他

除了上述介绍的方法之外,也有不少学者研究如何利用其他经典的分类方法来处理不平衡数据。其中包括神经网络[48,49]、K-近邻[50]以及模糊规则[51]等。值得一提的是,最近一种新颖的 minimax 算法被提出用于 IDS 分类[52,53]。其直接优化如下的最小最大化问题而获得分类超平面 $w^{\mathrm{T}}z=b$:

$$\max_{\alpha,\beta,b,w\neq 0}\ \alpha$$
$$\text{s. t.}\ \inf_{x\sim(\bar{x},\Sigma_x)}\Pr\{w^{\mathrm{T}}x\geqslant b\}\geqslant\alpha$$
$$\inf_{y\sim(\bar{y},\Sigma_y)}\Pr\{w^{\mathrm{T}}y\leqslant b\}\geqslant\alpha$$

其中:$x\sim(\bar{x},\Sigma_x)$表示总体 x 的均值为 $\bar{x}$ 和协方差 Σ_x,而不管其具体分布是什么。显然,从优化问题的形式看,minimax 算法实质是最大化最坏情形下的预测精度。其直观意义非常明确,而且由于它不涉及具体的分布形式,更具有稳健性。

6 结束语

不平衡数据 IDS 在实际应用中经常碰到,它对传统的分类方法构成了挑战。如何有效地处理 IDS 引起了人们的关注。IDS 分类也成了机器学习领域的又一新的研究热点[54,55]。目前,对 IDS 分类学习的研究主要集中在数据重抽样技术以及算法的改进方面。数据重抽样技术主要包括过抽样和欠抽样两种,它们各有优缺点;如何更合理地对数据进行重抽样是一个值得进一步研究的课题。在算法设计和改进方面,目前研究得比较多的是基于 SVM 的分类方法。SVM 的间隔最大化思想可以使分类器获得更好的推广能力,而且,真正决定 SVM 分类器的是那些只占少部分的支持向量样本。因此,与其他方法比较起来,SVM 方法似乎更适合处理不平衡数据。Boosting 技术与 SVM 有相似之处,其实质也是间隔最大化,因此可以尽可能地避免过学习所带来的不良影响;再者,Boosting 的适用范围更广,它可以提升各种弱分类算法。将 Boosting 用于 IDS 分类学习具有广阔的应用前景。

参考文献:

[1] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning, 1998,30(2-3):195-215.

[2] PHUA C, ALAHAKOON D. Minority report in fraud detection: classication of skewed data[J]. SIGKDD Explorations, 2004,6(1): 50-59.

[3] PÉREZ J M, MUGUERZA J, ARBELAITZ O, *et al*. Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance[C]//Proc of the 3rd International Conference on Advances in Pattern Recognition(ICAPR'05). 2005:381-389.

[4] CASTILLO M D del, SERRANO J I. A multistrategy approach for digital text categorization from imbalanced documents[J]. SIGKDD Explorations, 2004,6(1):70-79.

[5] ZHENG Zhao-hui, WU X, SRIHARI R K. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004,6(1):80-89.

[6] COHEN G, HILARIO M, SAX H, *et al*. Data imbalance in surveillance of nosocomial infections[C]//Proc of the 4th International Symposium on Medical Data Analysis(ISMDA'03). Berlin:[s. n.], 2003:109-117.

[7] CHEN Jian-xun, CHENG T H, CHAN A L F, *et al*. An application of classification analysis for skewed class distribution in therapeutic drug monitoring the case of vancomycin[C]//Proc of Workshop on Medical Information Systems(IDEAS-DH'04). Beijing:[s. n.], 2004:35-39.

[8] YOON K, KWEK S. An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics[C]//Proc of the 5th International Conference on Hybrid Intelligent Systems(HIS'05). Rio de Janeiro:[s. n.], 2005:303-308.

[9] RADIVOJAC P, KORAD U, SIVALINGAM K M, *et al*. Learning from class-imbalanced data in wireless sensor networks[C]//Proc of Vehicular Technology Conference(VTC'03-Fall). Orlando:[s. n.], 2003:3030-3034.

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

[11] JO T, JAPKOWICZ N. Class imbalances versus small disjuncts[J]. SIGKDD Explorations, 2004,6(1):40-49.

[12] PRATI R C, BATISTA G E A P A, MONARD M C. Learning with class skews and small disjuncts[C]//Proc of the 17th Brazilian Symposium on Artificial Intelligence(SBIA'04). Sao Luis:[s. n.], 2004:296-306.

[13] MERZ C J, MURPHY P M. UCI repository of machine learning databases[EB/OL]. (1999). http://www. ics. uci. edu/mlearn/MLRepository. html.

[14] PRATI R C, BATISTA G E A P A, MONARD M C. Class imbalances versus class overlapping: an analysis of a learning system behavior[C]//Proc of the 3rd Mexican International Conference on Artificial Intelligence(MICAI'04). Mexico City:[s. n.], 2004:312-321.

[15] FAWCETT T. ROC graphs: notes and practical considerations for researchers[EB/OL]. (2003). http://www. hpl. hp. com/personal/Tom Fawcett/papers/index. html.

[16] BREFELD U, SCHEFFER T. AUC maximizing support vector learning[C]//Proc of ICML Workshop on ROC Analysis in Machine Learning. Bonn:[s. n.], 2005.

[17] DRUMMOND C, HOLTE R C. Cost curves: an improved method for visualizing classifier performance[J]. Machine Learning, 2006,65(1):95-130.

[18] DASKALAKIL S, KOPANAS I, AVOURIS N. Evaluation of classifiers for an uneven class distribution problem[J]. Applied Artificial Intelligence, 2006,20(5):381-417.

[19] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. SIGKDD Explorations, 2004,6(1):20-29.

[20] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proc of 14th International Conference on Machine Learning(ICML'97). Nashville:[s. n.], 1997:179-186.

[21] BARANDELA R, VALDOVINOS R M, SÁNCHEZ J S, *et al*. The imbalanced training sample problem: under or over sampling[C]//Proc of International Workshops on Structural, Syntactic, and Statistical Pattern Recognition(SSPR/SPR'04). Lisbon:[s. n.], 2004: 806-814.

[22] DEHMESHKI J, KARAKÖY M, CASIQUE M V. A rule-based scheme for filtering examples from majority class in an imbalanced

training set[C]//Proc of MLDM 2003. 2003:215-223.

[23] CHAWLA N V, HALL L O, BOWYER K W, *et al*. SMOTE: synthetic minority oversampling technique[J]. Journal of Articial Intelligence Research, 2002,16:321-357.

[24] HAN Hui, WANG Wen-yuan, MAO Bing-huan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// Proc of International Conference on Intelligent Computing(ICIC'05). Hefei:[s. n.], 2005:878-887.

[25] ELKAN C. The foundations of cost-sensitive learning[C]//Proc of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01). Washington DC:[s. n.], 2001:973-978.

[26] DOMINGOS P. MetaCost: a general method for making classifiers cost-sensitive[C]//Proc of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'99). San Diego:[s. n.], 1999:155-164.

[27] FAN Wei, STOLFO S J, ZHANG Jun-xin, *et al*. AdaCost: misclassification cost-sensitive boosting[C]//Proc of the 16th International Conference on Machine Learning(ICML'99). Bled:[s. n.], 1999: 97-105.

[28] WEISS G. Mining with rarity: an unifying framework[J]. SIGKDD Explorations, 2004,6(1):7-19.

[29] RASKUTTI B, KOWALCZYK A. Extreme rebalancing for SVMs: a case study[J]. SIGKDD Explorations, 2004,6(1):60-69.

[30] AKBANI R, KWEK S, JAPKOWICZ N. Applying support vector machines to imbalanced datasets[C]//Proc of the 15th European Conference on Machine Learning(ECML'04). Pisa:[s. n,], 2004:39-50.

[31] WU Gang, CHANG E Y. KBA: kernel boundary alignment considering imbalanced data distribution[J]. IEEE Trans on Knowledge and Data Engineering, 2005,17(6):786-795.

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

[33] CHEN Xue-wen, GERLACH B, CASASENT D. Pruning support vectors for imbalanced data classification[C]//Proc of International Joint Conference on Neural Networks. Montreal:[s. n.], 2005:1883-1888.

[34] CALLUT J, DUPONT P. F_β support vector machines[C]//Proc of International Joint Conference on Neural Networks. Montreal:[s. n.], 2005.

[35] SCHÖLKOPF B, PLATT J C, SHAWE-TAYLOR J, *et al*. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001,13(7):1443-1472.

[36] MANEVITZ L M, YOUSEF M. One-class SVMs for document classication[J]. Journal of Machine Learning Research, 2001,2(1): 139-154.

[37] SENF A, CHEN Xue-wen, ZHANG A. Comparison of one-class SVM and two-class SVM for fold recognition[C]//Proc of ICONIP. Hong Kong:[s. n.], 2006:140-149.

[38] COHEN G, HILARIO M, PELLEGRINI C. One-class support vector machines with a conformal kernel: a case study in handling class imbalance[C]//Proc of International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR'04). Lisbon: [s. n.], 2004:850-858.

[39] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997,55(1):119-139.

[40] SCHAPIRE R E, FREUND Y, BARTLETT P, *et al*. Boosting the margin: a new explanation for the effectiveness of voting methods[J]. The Annals of Statistics, 1998,26(5):1651-1686.

[41] DEMIRIZ A, BENNETT K P, SHAWE-TAYLOR J. Linear programming boosting via column generation[J]. Machine Learning, 2002, 46(1-3):225-254.

[42] LESKOVEC J, SHAWE-TAYLOR J. Linear programming boosting for uneven datasets[C]//Proc of the 20th International Conference on Machine Learning. Washington D C:[s. n.], 2003:456-463.

[43] JOSHI M, KUMAR V, AGARWAL R. Evaluating boosting algorithms to classify rare classes: comparison and improvements[C]//Proc of the 1st IEEE International Conference on Data Mining. San Jose:[s. n.], 2001:257-264.

[44] CHAWLA N V, LAZAREVIC A, HALL L O, *et al*. SMOTEBoost: improving prediction of the minority class in boosting: knowledge discovery in databases[C]//Proc of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD'03). Cavtat Dubrovnik:[s. n.], 2003:107-119.

[45] GUO Hong-yu, VIKTOR H L. Learning from imbalanced data sets with boosting and data generation: the dataBoost-IM approach[J]. SIGKDD Explorations, 2004,6(1):30-39.

[46] LIU Yang, AN Ai-jun, HUANG Xiang-ji. Boosting prediction accuracy on imbalanced datasets with SVM ensembles[C]//Proc of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Singapore:[s. n.], 2006:107-118.

[47] KANG P, CHO S. EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems[C]//Proc of ICONIP. Hong Kong:[s. n.], 2006:837-846.

[48] ZHOU Zhi-hua, LIU Xu-ying. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2006,18(1):63-77.

[49] MURPHEY Y L, GUO H, FELDKAMP L A. Neural learning from unbalanced data[J]. Applied Intelligence, 2004,21(2):117-128.

[50] HAND D J, VINCIOTTI V. Choosing k for two-class nearest neighbour classifiers with unbalanced classes[J]. Pattern Recognition Letters, 2003,24(9-10):1555-1562.

[51] VISA S, RALESCU A. Learning imbalanced and overlapping classes using fuzzy sets[C]//Proc of ICML Workshop on Learning from Imbalanced Data Sets. 2003.

[52] LANCKRIET G, GHAOUI L, BHATTACHARYYA C, *et al*. A robust minimax approach to classification[J]. Journal of Machine Learning Research, 2003,3(1):555-582.

[53] HUANG Kai-zhu, YANG Hai-qin, KING I, *et al*. Imbalanced learning with biased minimax probability machine[J]. IEEE Trans on System, Man, and Cybernetics, 2006,36(4):913-923.

[54] JAPKOWICZ N. Learning from imbalaced data sets: a comparison of various strategies[C]//AAAI Workshop on Learning from Imbalanced Data Sets. Menlo Park:AAAI Press, 2000.

[55] CHAWLA N, JAPKOWICZ N, KOKCZ A. Editorial:Special issues on learning from imbalanced data sets[J]. SIGKDD Explorations, 2004,6(1):1-6.