不平衡数据分类方法综述

杨 明, 尹军梅, 吉根林

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 分类问题是机器学习领域的重要研究内容之一, 现有的一些分类方法都已经相对成熟, 用它们来对平衡数据进行分类一般都能取得较好的分类性能, 但在现实世界中数据往往都是不平衡的, 而现有的分类器的设计都是基于类分布大致平衡这一假设的, 如果用这些方法来对不平衡数据进行分类就会导致分类器的性能下降, 因而研究用于处理不平衡数据集的分类方法显得相当重要. 为便于读者更清晰地了解数据不平衡分类问题的研究现状和未来研究的动向, 本文对相关的研究进行了综述和展望.

[关键词] 不平衡数据, 过抽样, 欠抽样, 代价敏感, 单分类器, 特征选择, 子空间

[中图分类号] TP 311 [文献标识码] A [文章编号] 1672-1292(2008)04-0007-06

Classification Methods on Imbalanced Data: a Survey

Yang Ming, Yin Junmei, Ji Genlin

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract Classification is one of the most important research contents in machine learning, and the traditional classification methods are relatively mature, when dealing with well-balanced data they can make good performance. But in real world the data is usually imbalanced. The design of the existing classification methods is often based on the assumption that the training sets are well-balanced, so it may lead to the descending capability of the classification methods when dealing with imbalanced data. Making researches on imbalanced data is quite important. In order to help readers to have a clear idea of the currently proposed and future work on the issue of unbalanced data classification, we make a simple survey of the studies of this issue and give some key problems attracting researchers in this paper.

Key words imbalanced data, over-sampling, under-sampling, cost-sensitive, one classifier, feature selection, subspace

分类问题是机器学习领域的重要研究内容之一, 现有的一些分类方法都已经相对成熟, 用它们来对平衡数据进行分类一般都能取得较好的分类性能. 然而现有的分类器的设计都是基于类分布大致平衡这一假设的, 通常假定用于训练的数据集是平衡的, 即各类所含的样本数大致相当, 然而这一假设在很多现实问题中是不成立的, 数据集中某个类别的样本数可能会远远少于其它类别.

许多实际的应用领域中都存在不均衡数据集, 例如欺骗信用卡检测、医疗诊断、信息检索、文本分类等, 其中少数类的识别率更为重要. 在医疗诊断中如果把正常人误诊为病人固然会给他带来精神上的负担, 但如果把一个病人误诊为正常, 就可能会错过最佳治疗时期, 从而造成严重的后果. 传统的分类方法倾向于对多数类有较高的识别率, 对于少数类的识别率却很低. 因此不均衡数据集的分类问题的研究需要寻求新的分类方法和判别准则.

鉴于解决不平衡学习问题有着很深远的意义, 因此研究者对该问题进行了大量的研究. 相关研究主要围绕以下 3个方面展开: (1) 改变数据的分布; (2) 设计新的分类方法; (3) 设计新的分类器性能评价准则. 为便于读者更加清晰地了解针对数据不平衡分类问题的研究现状和未来的研究动向, 本文对此做一个概要性介绍并进行了展望.

收稿日期: 2008-06-18
基金项目: 国家自然科学基金(60873176)资助项目.
通讯联系人: 杨 明, 教授, 博士, 研究方向: 数据挖掘、机器学习和粗糙集理论及应用. E-mail: myang@njnu.edu.cn

1 不平衡数据分类问题的难点

不同于均衡数据的分类, 不平衡数据的分类问题求解相对较难, 其主要原因如下:

(1) 经典的分类精度评价准则不能适用于不平衡数据的分类器性能判别. 在传统机器学习中通常采用分类精度作为评价准则, 当对不平衡数据进行学习时, 少数类对分类精度的影响可能会远远小于多数类. W eiss实验研究表明, 以分类精度为准则的分类学习通常会导致少数类样本的识别率较低[1], 这样的分类器倾向于把一个样本预测为多数类样本. 若训练数据是极端不平衡的, 学习的结果可能没有针对少数类的分类规则, 因此对于不平衡数据的分类, 以高分类精度为目标是不合适的, 需要引入更加合理的评价标准[2].

(2) 仅有很少的少数类样本数据[3-5]. 仅有很少的少数类样本分两种情况: 少数类样本绝对缺乏和少数类样本相对缺乏. 无论哪种情况, 我们称类分布的不平衡程度为少数类中的样本数与支撑类中的样本数之比. 在实际应用中, 该比例可以达到 1∶100, 1∶1 000甚至更大[5]. 文献[1]对该比例与分类性能之间的关系进行了深入的研究, 研究结果表明很难明确地给出何种比例会降低分类器的性能, 这是因为分类器的性能还与样本数和样本的可分性有关. 在某些应用下, 1∶35的比例就会使某些分类方法无效, 甚至 1∶10的比例也会使某些分类方法无效.

对情况 1, 因少数类所包含的信息就会很有限, 从而难以确定少数类数据的分布, 即在其内部难以发现规律, 进而造成少数类的识别率低. 已有研究表明分类器的误差率大多源于小析取项[3]. 事实上小析取项本质上并非比大析取项更容易产生错误, 使小析取项更容易产生错误的原因有分类器的偏移、噪声的干扰和缺失属性等. 为有效避免因噪声信息的影响, 如何判别有意义的小析取项是值得研究的课题[3, 4].

对情况 2, 少数类样本数据相对缺乏不同于少数类样本数据的绝对缺乏, 相对缺乏是指少数类样本在绝对数量上并不少, 但相对于多数类来说它的样本数目很少. 在样本相对缺少的情况下, 同样不利于少数类的判别, 这是因为多数类样本会模糊少数类样本的边界, 且使用贪心搜索法难以把少数类样本与多数类区分开来, 而更全局性的方法通常难以处理[5].

(3) 数据碎片. 从算法设计角度看, 很多分类算法采用分治法, 这些算法将原始的问题逐渐分为越来越小的一系列子问题, 因而导致原空间被划分为越来越小的一系列子空间, 样本空间的逐渐划分会导致数据碎片问题, 这样只能在各个独立的子空间中寻找数据的规律, 对于少数类来说每个子空间中包含了很少的数据信息, 一些跨空间的数据规律就不能被挖掘出来[5]. 数据碎片问题也是影响少数类样本学习的一个突出的问题.

(4) 不恰当的归纳偏置. 根据特定样本的归纳需要一个合理的偏置[3], 否则学习就不能实现. 归纳偏置对算法的性能有着很大的影响, 为了获得较好的性能并避免过度拟合, 许多学习算法使用的偏置往往不利于对少数类样本的学习. 许多归纳推理系统在存在不确定时往往倾向于把样本分类为多数类. 可见, 不恰当的归纳偏置对不平衡数据的学习是不利的.

此外, 大多数分类器的性能都会受噪声的影响. 在不平衡问题中, 由于少数类的数量很少, 因此分类器有可能难以正确区分少数类和噪声, 故噪声对少数类的影响要大于对多数类的影响. 噪声的存在使防止过拟合技术变得非常重要, 如何抑制噪声、强化少数类样本的作用是具有挑战性的研究工作.

2 不平衡数据分类的相关方法

目前不平衡数据分类的相关方法主要从数据层面、算法层面和判别准则 3个不同层面进行研究, 本节主要从数据层面和算法层面上对不平衡数据分类的相关方法进行概要性的回顾和综述, 基于新的判别准则的不平衡分类问题求解方法将在第 3部分进行介绍.

从数据层面采用的策略来看, 大体上采用过抽样、欠抽样两种方法. 具有描述如下:

(1) 过抽样. 抽样处理不平衡数据的最常用方法, 基本思想就是通过改变训练数据的分布来消除或减小数据的不平衡. 过抽样方法通过增加少数类样本来提高少数类的分类性能[6], 最简单的办法是简单复制少数类样本, 缺点是引入了额外的训练数据, 会延长构建分类器所需要的时间, 没有给少数类增加任何新的信息, 而且可能会导致过度拟合. 改进的过抽样方法通过在少数类中加入随机高斯噪声或产生新的合

成样本等方法, 一定程度上解决了上述问题, 如 Chawla等人提出的 SMOTE算法[7]. 此外, 文献[8]提出了一种基于初分类的过抽样算法, 基本思想是: 一个多数类样本, 若它在训练集中的 k个近邻也都属于多数类, 根据 k近邻的思想则该样本离分类边界较远, 对分类是相对安全的. 将多数类中满足上述条件的所有样本放入集合 E, 将少数类与集合 E 合并记为训练集 A, 利用 A 对多数类样本进行最近邻分类, 误分类的多数类样本放入集合 H. 将少数类和集合 H 合并为第二个新的训练集 B.

(2) 欠抽样. 欠抽样方法通过减少多数类样本来提高少数类的分类性能, 最简单的方法是通过随机地去掉一些多数类样本来减小多数类的规模, 缺点是会丢失多数类的一些重要信息, 不能够充分利用已有的信息[9]. 因此人们提出了许多改进的欠抽样方法. Kubat等人提出的 One-sided selection算法尽可能地不删除有用的样本, 多数类样本被分为"噪音样本"、"边界样本"和"安全样本", 将边界样本和噪音样本从多数类中删除, 得到的分类效果会比随机欠抽样理想一些. 也可以把对少数类的过抽样与对多数类的欠抽样两者结合起来[7]. One-sided selection算法是通过判断样本间的距离的方式来把多数类划分为"噪音样本"、"边界样本"和"安全样本"的, 文献[10]提出了一种基于 genetic algorithms (GA)的方式来对多数类进行抽样, 找出噪音样本并将它们去除. 文献[11]提出了一种基于聚类的欠抽样算法, 先用聚类的方法将训练集划分成几个簇, 每个簇都包含一定数目的多数类和少数类, 对每个簇, 取出其中所有的少数类, 然后按照一定规则对该簇中的多数类进行欠抽样, 最后将从每个簇中取出的样本进行合并, 得到一个新的训练集, 对其进行训练.

从算法层面上看, 采用的相关策略如下:

(1) 代价敏感方法. 在处理不平衡问题时, 传统的分类器对少数类的识别率很低, 对多数类的识别率却很高, 然而在现实生活中往往是少数类的识别率更为重要, 因此少数类的错分代价要远远大于多数类. 例如在入侵检测中, 可能在 1 000次通信中只有少数几次是攻击, 但将攻击误报为正常和将正常误报为攻击所引起的代价是截然不同的. 在代价敏感学习方法中, 代价信息通常由领域专家给出, 在进行学习时假设各个类别的代价信息是已知的, 在整个学习过程中是固定不变的.

目前对代价敏感学习的研究主要集中在以下两个方面: ① 根据样本的不同错分代价重构训练集, 不改变已有的学习算法. 重构训练集的方法是根据样本的不同错分代价给训练集中的每一个样本加权, 接着按权重对原始样本集进行重构, 但其存在的缺点就是重构的过程中丢失了一些有用样本的信息. ② 在传统的分类算法的基础上引入代价敏感因子, 设计出代价敏感的分类算法. 代价敏感的学习中不同类的错分代价是不同的, 通常多数类的代价比少数类大得多, 对小样本赋予较高的代价, 大样本赋予较小的代价, 期望以此来平衡样本之间的数目差异[12]. 文献[13]对 Veropoulos的代价敏感 SVM 进行了改进, 但基本思想都是将代价与松弛变量相关联来使 SVM 的超平面对代价敏感. Lee等人为多类问题设计了代价敏感的 SVM, 并考虑了采样偏置[14]. 文献[15]提出了一种加权 Fisher线性判别模型 (WFLD), 通过对正负两类的类内离散度矩阵进行分别加权, 使正负两类样本的协方差矩阵对总类内离散度矩阵的贡献平衡. 但这些代价敏感学习方法主要是针对全局模型. 为此, 文献[16]提出了一种局部代价敏感算法, 预测一个新样本时, 首先选定一种合适的距离度量方式, 选出该测试样本的 k个近邻, 然后用加权的方式对这 k个近邻进行训练, 得到一个分类器用来进行预测.

(2) 集成学习方法. 按照基本分类器之间的种类关系可以把集成学习方法划分为异态集成学习和同态集成学习两种[17], 异态集成学习指的是使用各种不同的分类器进行集成, 同态集成学习是指集成的基本分类器都是同一种分类器, 只是这些基本分类器之间的参数有所不同. 在不平衡数据的分类问题上, 由于异态集成学习的每个基本算法都有独到之处, 因而某种基本算法会对某类特定数据样本比其余的基本算法更为有效. 同态集成学习方法中针对不平衡数据的多数是把抽样与集成结合起来, 对原始训练集进行一系列抽样, 产生多个分类器, 然后用投票或合并的方式输出最终结果.

AdaBoost应用于不平衡数据分类可取得较好效果, 但有实验结果表明 AdaBoost提高正类样本的识别率的能力有限[18], 因为 AdaBoost是以整体分类精度为目标的, 负类样本由于数目多所以对精度的贡献大, 而正类样本由于数目很少因此贡献相当小, 故分类决策是不利于正类的. 为此, 一些改进相继被提出, 如 AdaCost[19]、RareBoost[20], 主要策略是改变权值更新规则, 使分类错误的正类样本比负类样本有更高的权值. 文献[21, 22]将过抽样与集成方法进行融合, 既能利用过抽样的优点增加少数类样本的数量, 使分类

器能够更好地提高少数类的分类性能,又能利用集成方法的优点提高不平衡数据集的整体分类性能,如文献[23]提出的 C-SMOTE算法就是过抽样和集成结合的成功例子.

(3) 单类分类器方法. 在实际应用中有时想要获取两类或多类样本是很困难的,或者就是需要很高的成本,只能获取单类样本集. 在这种情况下,对只含有单一类的数据进行训练是唯一可能的解决办法. 单分类器是用来对只有一种类别的训练集进行分类的,它是一个能有效解决不平衡数据问题的办法[3]. 文献[13]用 SVM 来仅对正类进行训练,实验表明该方法是有效的. 单类分类器由于只需要一类数据集作为训练样本,训练数据量变小了,从而减少了构建分类器所需要的时间,节约了开销,因此在很多领域都有着良好的应用前景.

(4) 面向单个正类的 FLDA方法. 某些极端的情况下,少数类只有一个样本. 针对该情况,文献[24]提出了一种解决单个正例的 Fisher线性判别方法,找出单个正例在负类中的 k个近邻,然后按照一定规则依次在单个正例和它的各个近邻的连线上产生合成样本,并把这些合成样本添加到原始的正类中,再用分类器进行分类.

(5) 多类数据不平衡问题的解决方法. 目前针对多类的不平衡问题研究较少,通常仍然采用已有的多类分类方法和两类不平衡分类策略结合. Lee等人结合采样偏置[14],提出针对多类的代价敏感的 SVM. 文献[25]用 k个超球来对 k类数据进行描述,其中每个超球包含一类数据. Wang Defeng等人提出了结构性单分类器(Structured One-Class Classification),考虑了数据分布情况,将一类目标数据用多个超椭球来描述,通过对目标类的描述将目标数据与非目标数据分开. 文献[3]在 Structured One-Class Classification的基础之上对多类样本中的每一类都用多个超球描述,而不仅将一类目标数据用多个超球来描述,该算法根据每一类样本的数目多少用多个超球来包围起来,以获得比单个超球更好的描述.

(6) 其它方法. 主动学习[27]、随机森林[28]、子空间方法[29]、特征选择方法[30,31]和 SVM 模型下的后验概率求解方法[32]等也是学习不平衡数据集的有效方法.

3 不平衡数据分类的评价准则

精确度 $\text{accuracy} = \dfrac{(TP + TN)}{(TP + TN + FP + FN)}$是分类问题中常用的评价标准(如表 1所示),它反映分类器对数据集的整体分类性能,但不能正确反映不平衡数据集的分类性能. 为此,针对不平衡数据,需提出更为合理的评价标准,常用的标准有:查全率 recall,查准率 precision、F-value值、G-mean值、AUC. 少数类的 recall、precision、F-value、G-mean值的计算方法分别如下:

$$\text{Recall} = TP/(TP + FN); \tag{1}$$

$$\text{Precision} = TP/(TP + FP); \tag{2}$$

$$\text{F-value} = \left(\frac{(1+\beta^2)*\text{recall}*\text{precision}}{(\beta^2*\text{recall}+\text{precison})}\right); \tag{3}$$

$$\text{G-mean} = \sqrt{\frac{TP}{TP + FN}*\frac{TN}{TN + FP}}. \tag{4}$$

少数类的 F-value是不平衡数据集学习中有效的评价准则,它是 Recall和 Precision的组合,其中 β是可调参数,通常取值为 1. 仅当少数类的 Recall和 Precision的值都大时,它的 F-value才会大,因此它能正确地反映少数类的分类性能. 此外 G-mean也是不平衡数据集学习中常用的评价准则,它是少数类的精确度 $TP/(TP + FN)$与多数类的精确度 $TN/(TN + FP)$的乘积的平方根,二者的值都大时,G-mean才会大,因此 G-mean能合理地评价不均衡数据集的总体分类性能.

研究表明 ROC(Receiver Operating Characteristic)曲线能够全面地描述分类器在不同判决阈值时的性能,已成为不平衡数据分类器性能评价的准则. 让 $x = FP/(TN + FP)$, $y = TP/(TP + FN)$,将 x和 y分别作为横、纵坐标. 每一个阈值对应一个 (x, y)点,改变阈值,将得到的所有 (x, y)点连起来就是分类器在该数据集上的 ROC曲线,ROC

表 1 混合矩阵

Table 1 Confusion matrix

	被分为正类	被分为负类
实际为正类	*TP*	*FN*
实际为负类	*FP*	*TN*

曲线越靠近左上角表示分类器性能越好. 然而, ROC曲线不能定量的对分类器的性能进行评价, 于是人们常常采用 ROC曲线下面积 AUC来代替 ROC曲线对分类器的性能进行评估[33, 34], AUC值越大则分类器的性能就越好.

4　结语

不平衡数据的存在是妨碍机器学习被广泛使用的一个重要原因, 近年来这个问题引起了广泛关注. 不平衡问题普遍存在于许多实际应用领域中, 其中研究者特别关注少数类的分类性能的提高, 如何有效地提高它的分类性能是研究者追求的共同目标. 针对数据不平衡分类问题, 人们提出了很多的解决方法, 且取得了一定的进展, 但仍有很多问题需要进行深入研究, 如: 关于算法的效率和时间开销方面研究、如何自适应地确定最好的抽样比例等. 目前绝大多数的不平衡问题的研究都是针对数据数目比例失衡的情况来考虑的, 不平衡数据还有另外一种情况, 就是两类数据数目相当, 但是类分布差别较大, 一类比较集中, 另一类比较分散, 目前关于类分布差异的研究较少. 此外, 如何将特征选择方法融入到不平衡分类算法中也是今后需要进一步研究的问题.

[参考文献] (References)

[1] Weiss G M, Provost F. Learning when training data are costly: the effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, 19: 315-354.

[2] Zadrozny B, Elkan C. Learning and making decisions when costs and probabilities are both unknown[C] // Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001: 204-213.

[3] 缪志敏. 基于单分类器的数据不平衡问题研究[D]. 南京: 中国人民解放军理工大学指挥自动化学院, 2008.
Miao Zhimin. Research on imbalanced data based on one-class classifiers[D]. Nanjing: Institute of Automation Command, PLA University of Science and Technology, 2008. (in Chinese)

[4] Holte R C, Acker L E, Porter B W. Concept learning and the problem of small disjuncts[C] // Proceedings of the 11th International Joint Conference on Artificial Intelligence. Austin: Morgan Kaufmann, 1989: 813-818.

[5] Sun Y M, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalance data[J]. Pattern Recognition, 2007, 40: 3358-3378.

[6] Maloof M A. Learning when data sets are imbalanced and when costs are unequal and unknown[C] // ICML-2003 Workshop on Learning from Imbalanced Data Sets II. Washington DC: AAAI Press, 2003.

[7] Chawla N, Bowyer K, Hall L, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[8] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans Knowl Data Eng, 2006, 18(1): 63-77.

[9] Weiss G M. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations, 2004, 6(1): 7-19.

[10] Drown D J, Khoshgoftaar T M, Narayanan R. Using evolutionary sampling to mine imbalanced data[C] // The 6th International Conference on Machine Learning and Applications. Washington DC: IEEE Computer Society, 2007: 363-368.

[11] Yen S J, Lee Y S. Cluster-based under-sampling approaches for imbalanced data distributions[C] // Proceedings of the 8th International Conference. Berlin: Springer, 2006: 427-436.

[12] Ciraco M, Rogalewski M, Weiss G. Improving classifier utility by altering the misclassification cost ratio[C] // Proceedings of the 1st International Workshop on Utility-based Data Mining. New York: ACM, 2005: 46-52.

[13] Raskutti B, Kowalczyk A. Extreme rebalancing for SVMs: a case study[J]. News letter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004, 6(1): 61-69.

[14] Lee Y, Lin Y, Wahba G. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data[R]. Wisconsin: University of Wisconsin, 2002.

[15] 谢纪刚, 裘正定. 不平衡数据集 Fisher线性判别模型[J]. 北京交通大学学报, 2006, 30(5): 15-18.
Xie Jigang, Qiu Zhengding. Fisher linear discriminant model with class imbalance[J]. Journal of Beijing Jiaotong University, 2006, 30(5): 15-18. (in Chinese)

[16] Karagiannopoulos M G, Anyfantis D S, Kotsiantis S B, et al. Local cost sensitive learning for handling imbalanced data sets[C] // 2007 Mediterranean Conference on Control and Automation. Athens: IEEE Press, 2007: 1-6.

[17] Yu Shixin. Feature selection and classifier ensembles: a study on hyperspectral remote sensing data[D]. Flanders: Universi-

ty of Antwerp, 2003.
[18] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning, 1999, 37(3): 297-336.
[19] Fan W, Stolfo S J, Zhang J, et al. AdaCost: misclassification cost-sensitive boosting[C]// Bratko I, Dzeroski S. Proc of the 16th Intern Conf on Meachine Learning. Morgan Kaufmann, 1999: 97-105.
[20] Joshi M V, Kumar V, Agarwal R C. Evaluating boosting algorithms to classify rare classes: comparison and improvements [C]// Cercone N, Lin T Y, Wu X. Pro of the 2001 IEEE Intern Conf on Data Mining. Washington DC: IEEE Computer Society Press, 2001: 257-264.
[21] Chawla N V, Japkowicz, Kolcz A. Editorial: special issue on learning from imbalanced data sets[J]. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 1-6.
[22] Chawla N V, Lazarevic A, Hall L O. SMOTEBoost: improving prediction of the minority class in boosting[C]// The 7th European Conf on Principles and Practice of Knowledge Discovery in Databases. Berlin: Springer, 2003: 107-119.
[23] He Guoxun, Han Hu, Wang Wenyuan. An over-sampling expert system for learning from imbalanced data sets[J]. Neural Networks and Brain, 2005, 1: 537-541.
[24] 尹军梅,杨明. 一种面向单个正例的Fisher线性判别分类方法[J]. 南京师范大学学报:工程技术版, 2008, 8(3): 61-65.
Yin Junmei, Yang Ming. A fisher discriminant classification approach dealing with single positive sample[J]. Journal of Nanjing Normal University: Engineering and Technology Edition, 2008, 8(3): 61-65. (in Chinese)
[25] Tao Ban, Shigeo Abe. Implementing multi-class classifiers by one-class classification methods[C]// 2006 International Joint Conference on Neural Networks. Sheraton Vancouver Wall Centre Hotel, Vancouver, BC: IEEE Press, 2006: 16-21, 327-332.
[26] Sun Y. Cost-sensitive boosting for classification of imbalanced data[D]. Canada: University of Waterloo, 2007.
[27] Constantinopoulos C, Likas A. Semi-supervised and active learning with the probabilistic RBF classifier[J]. Artificial Neural Networks, 2008, 71(13): 2489-2498.
[28] Chen C, Liaw A, Breiman L. Using random forests to learn unbalanced data[R]. California: University of California, 2004.
[29] Ahn H, Moon H, Fazzari M J, et al. Classification by ensembles from random partitions of high-dimensional data[J]. Computational Statistics & Data Analysis, 2007, 51: 6166-6177.
[30] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004, 6(1): 80-89.
[31] Mladenic D, Grobenik M. Feature selection for unbalanced class distribution and Naïve Bayes[C]// Proceedings of the 16th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1999: 258-267.
[32] Tao Q, Wu G, Wang F Y, et al. Posterior probability support vector machines for unbalanced data[J]. IEEE Trans on Neural Networks, 2005, 16(6): 1561-1573.
[33] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(7): 1145-1159.
[34] Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers, HPL-2003-4[R/OL]. [2008-06-18] http://www.purl.org/NET/tfawcett/papers/ROC101.pdf

[责任编辑:严海琳]