不平衡类数据挖掘研究综述

翟 云[1,2] 杨炳儒[1] 曲 武[1]
(北京科技大学信息工程学院 北京 100083)[1] (聊城大学计算机学院 聊城 252059)[2]

摘 要 综述了近年来国内外对不平衡类数据挖掘的主要研究进展。首先分析了不平衡类数据挖掘的本质。其次,详细探讨了处理不平衡类数据挖掘的各种技术,并根据其本质区别,从数据层次和算法层次分别对目前存在的各种技术方法进行了深入剖析和全面比较。最后,指出当前不平衡类数据挖掘研究的热点以及将来需要重点关注的主要问题。

关键词 机器学习,不平衡类数据,重采样,代价敏感学习
中图法分类号 TP181 **文献标识码** A

Survey of Mining Imbalanced Datasets

ZHAI Yun[1,2] YANG Bing-ru[1] QU Wu[1]
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)[1]
(College of Computer Science, Liaocheng University, Liaocheng 252059, China)[2]

Abstract This paper reviewed the present situation of mining data in imbalanced classes at home and abroad in recent years. Firstly, it analysed in-depth the existing problems and their resulting nature. Then it in detail dealt with various state-of-the-art data mining techniques under the imbalanced learning scenario. Moreover, from the data-level and algorithm-level respectively it analysed and compared them comprehensively in accordance with essential difference. At the same time, the paper summaried measure metrics evaluating performance of mining imbalance data sets. Also, the paper pointed out recent hot issues of theoretic studies and applications. Finally, the perspectives on future work were also discussed.

Keywords Machine learning, Imbalanced classification, Resampling, Cost sensitive learning

在两分类数据集中,数量相当少的一类被称为少数类或稀缺类(minority class),而另一类则被称为多数类(majority class),具有这样特征的两分类数据集则被称为是不平衡的。正是由于少数类的样本和多数类的样本分别代表稀缺样本的存在与否,故它们通常分别被称为正样本(positive examples)和负样本(negative examples)。现实世界里,不平衡类问题是常见的,如通过对不同病人检查形成的一系列乳房-射线数据库已经在处理不平衡类数据算法中得到广泛应用。其中,癌变和健康的病例分别分到少数类和多数类。事实上,非癌变的病人数目要远远大于癌变的病人数目,在数据集中,存在10923个多数类样本和260个少数类样本。在其他应用如信用卡欺骗检测[1,2]、文本分类[3]、信息搜索及过滤[4]、市场行为分析[5]等中,人们主要关心的是数据集中的少数类,但这些少数类的错分所产生的代价异常大,甚至是不可估量的。因此,在实际应用中,通过数据挖掘技术提高少数类的分类精度进而减少由于误分类造成的重大损失的研究任务迫在眉睫。近几年来,不平衡数据集的分类问题开始受到数据挖掘界的重视。在2000年AAAI(the Association for the Advancement of Artificial Intelligence)国际会议和2003年ICML(the international Conference on Machine Learning workshop on Learning from Imbalanced Data Sets)国际会议上分别设立论坛进行专题研讨。2005年在ICDM(IEEE International Conference on Data Mining)国际会议上,不平衡类数据挖掘被列为数据挖掘领域的十大挑战性难题之一。

本文第1节介绍不平衡数据分类问题的本质,从不平衡类分布、样本规模、分离性(separability)以及类内子聚集(subcluster)等4个方面展开讨论。第2节、第3节分别从数据层次和算法层次对目前存在的各种技术方法进行剖析和比较。最后进行总结和回顾,并对未来的研究进行展望。

1 不平衡数据分类问题的本质

具有类不平衡问题的数据集的最主要特征就是各个类之间的数据不平衡分布。然而,理论和实验的研究[6-8]表明,在辨别稀缺样本时,类不平衡分布并不是影响分类器模型的唯一因素。其他的因素包括样本规模、分离性以及类内存子聚集。

1.1 不平衡类分布

在两分类情况下,类分布的不平衡度通常定义为:

$$\text{不平衡度 } D=\frac{\text{正类样本规模}}{\text{负类样本规模}} \quad (1)$$

在实际应用中,不平衡度可能为1∶100,也可能是1∶

到稿日期:2009-11-06 返修日期:2010-01-25 本文受国家自然科学基金(60675030,60875029)资助。

翟 云(1979-),男,博士生,讲师,CCF会员,主要研究方向为知识发现,E-mail:yunfei_2001_1@yahoo.com.cn;杨炳儒(1943-),男,教授,博士生导师,主要研究方向为知识发现与智能系统、柔性建模与集成技术;曲 武(1982-),男,博士生,主要研究方向为知识发现。

1000，甚至更大[9]。文献[10]研究了训练数据集的类分布与决策树分类器的分类性能之间的关系。结果显示，相对平衡的类分布通常会得到较理想的结果；样本规模、分散性等也影响着分类性能，因此并不能准确地描述出类不平衡达到什么程度时会导致分类性能下降。在某些应用中，当不平衡度为1∶50时可以构造一个较好的分类模型，然而取1∶10时却难以构造这样的分类模型。

1.2 样本规模

当不平衡度固定时，样本规模对决定分类模型的好坏起到了至关重要的作用。样本规模较小时，难以发现少数类样本内在的规则。随着训练集的增大，由不平衡类引起的错误率在不断减小[7]。其中的道理显而易见：采用更多的数据时，少数类中更多的有用信息将有助于分类模型的形成。只要数据集足够大，类不平衡分布也不可能对分类模型造成太大影响。

1.3 分离性

将少数类从数据集中正确分离是解决不平衡类问题的核心。假定每类都存在较高的可识别的模式，就不需要用太复杂的模式区分每一个样本。然而，如果每类的模式在特征空间中存在着不同程度的重叠，就难以推导出区分规则。当类间重叠度在变化时，类不平衡分布本身似乎并不是问题，但是当类高度重叠时，少数类正确分类的数目显著减少[7]。

1.4 类内子聚集

在许多分类问题中，每类数据集通常包括几个子聚集，或子概念。类样本来自不同子聚集，这些子聚集并不包含相同数目的样本，这种现象被称为类内不平衡。类内不平衡子聚集的存在进一步加剧了类不平衡分布问题，具体体现在以下3方面：

(1)类内不平衡子聚集的存在增加了数据集概念学习复杂性；

(2)多数情况下，类内不平衡子聚集难以描述；

(3)类内不平衡与类间不平衡两者难以有效结合。

2 重采样(resampling)技术

在数据层面解决非平衡数据分类问题的方法主要有两种：对少数类过采样(over-sampling)和对多数类欠采样(under-sampling)，直到两类数据趋于均衡。然而，事实已经表明，两种技术均有重大缺陷。欠采样有可能漏掉潜在重要的数据，而过采样人为地增加了数据数量，加重了学习算法的计算负担，而且有可能导致过适应(over-fitting)；但两种方法均因改变了原始数据的分布而备受争议。

2.1 过采样技术

增加少数类数据的最常用方式是随机过采样，也就是说用非启发式方法通过复制正样本使两类数据趋于平衡。此方法简单宜用，但由于反复复制了少数类数据而增加了过拟合的可能性。

Chawla 等提出了新的增加少数类的过采样方法。由于仅简单复制属于少数类的样本只能通过插入更多相邻的样本达到增加更多少数类样本数量的目的，Chawla 提出了SMOTE(Synthetic Minority Oversampling TEchnique)[11]智能过采样方法，它不仅仅是通过简单复制样本达到平衡，而是通过改进的采样技术，力图在训练集中添加有用信息。在SMOTE中，从已存在少数类样本集中产生新的样本的步骤如下：找到已存在样本点的 n(如常用5个)个最近邻居；从中随机选取一个；新的数据点是从连接原始数据点与随机选择的邻居的直线上选择的。这种技术不是简单地复制少数类数据，而是扩大了分类决策区域，从少数类数据集中复制更大范围的数据而到达数据平衡之目的。实验证明，SMOTE性能远远超出随机过采样技术[11,12]。

SMOTE算法进行了多次修改，力图生成更多正样本的区域。Han 等提出 Borderline-SMOTE 算法[13]来对 SMOTE 进行改进。Borderline-SMOTE 仅复制了紧靠特征空间决策边界的正样本，而这些样本恰恰是最容易被误分类的。同时，采用 SMOTE 算法对这些样本进行采样，而不是直接对它们或所有随机子集数据过采样。实验证明，Borderline-SMOTE 比 SMOTE 和随机过采样技术取得了更好的分类性能。Alexander Liu 等提出的产生式过采样(generative overs-ampling)技术[14]，从数据点集中学习得到新的数据点，在文本分类数据集中显示出良好性能。

尽管过采样技术增加了学习算法的计算代价，但 Batista 等通过实验证明，当数据集有非常少的正样本时，该技术简单方便[15]；Barandela 等在多数类与少数类比例较大时得到同样的结论[16]。

2.2 欠采样技术

欠采样技术包括随机欠采样和改进的欠采样算法。随机欠采样技术旨在通过随机取出部分负样本而达到平衡数据类的目的[17]。尽管简单，但通过实验已经证明该方法是有效的采样技术之一[18]。该技术的主要问题在于它可能丢弃一些分类过程中潜在的重要数据，而且分类器的机器学习目的是要估计目标样本的概率分布，而样本的概率分布通常是未知的。

与随机欠采样技术不同，许多新欠采样技术主要通过更加智能化的方式从多类数据中选择多类而达到数据集类间平衡之目的。Kubat 等提出了一种称作单边选择(one-sided selection)的新欠采样技术，其思想是通过去除一些认为是多余或噪音的数据而力图实现对多数类进行智能欠采样[19]。具体做法是：把多数类的数据分为3种类型：安全的(safe)、边界的(borderline)和噪音(noise)数据。选择性地去除冗余数据或毗邻少数类的数据(这些边界数据通常被认为是噪音数据)，通过 Tomek 连接[20]判断边界样本。

Barandela 等通过 Wilson's 算法来清洗数据，削减多数类的样本数[16]。与此类似，Batista 等削减了标签(label)与它们的3-近邻不一致的样本数，从而减小了稀缺类的误分类错误。上述各方法的主要缺点在于并不存在能够去除多数类模式的控制器，因此常常导致类分布不平衡。而文献[21]利用KNN技术对样本进行分类，并且去除了误分类的那些多数类中的样本；还提出了一种修正距离计算方案，即通过对样本偏置从而使其更被趋向分到正类。

其它的一些工作利用遗传算法来减少多数类数据集，直到类间数据基本平衡为止。Barandela 等通过类似实验瞬间选择属性，在不平衡区域去除部分多数类样本[11]。文献[22]提出了一种基于限制性"去污"(decontamination)的采样技术，该技术在在减少多数类的同时也增加了少数类的数量；这种技术在去掉了一些负样本的同时对部分其它样本重新标签

(relabel),从而使得多数类样本数量减少。同时,来自多数类的一些样本通过改变标签被重新组合到少数类中,使少数类数量增多。

值得说明的是一些上述研究已经证实当类间平衡性很差时采用欠采样技术简单有效。

2.3 混合采样技术

如上所述,单一的过采样技术和欠采样技术都不可避免地存在着缺陷。Ling 和 Li 把对少数类的过采样与对多数类的欠采样技术组合在一起,但是遗憾的是,两者的组合并没有使性能得到明显改善[23]。尽管如此,混合采样技术作为处理不平衡类数据的一项新技术而得到越来越多的关注。Cohen 等先利用特定类的子聚类算法对数据进行预处理,通过得到的矩心(centroids)定义新的样例,然后通过对多数类的欠采样和少数类进行过采样得到合成样本[10]。

2.4 特征选取

特征选择就是去除特征空间中一部分冗余或无用的特征,目的在于减小特征维数。特征选择技术在文本分类问题中得到广泛应用。文本分类是数据挖掘领域的研究热点之一。文本分类中数据集的不均衡问题是一个在实际应用中普遍存在的问题。样本数量分布很不平衡时,特征的分布同样会不平衡;在多数类中经常出现的特征,在少数类中很少出现或者根本不出现。根据不平衡分类问题的特点,选取最具有区分能力的特征,有利于提高少数类的分类精确率。因此,特征选择方法对于不平衡分类问题的研究同样具有非常重要的意义。

针对文本分类中存在的不平衡分类问题,Zheng 等认为用于特征选择已存在的性能指标不适用于非平衡类数据集[24]。他们提出了一种新的特征选择框架,即分别选取正类和负类的特征,然后再巧妙组合起来。结果显示,Bayes 技术利用该方法与一些较成熟的算法如逻辑回归等具有较强的竞争力。

2.5 高级采样方法

与前面的各种重采样方法不同,下面的重采样技术基于先前的分类结果。

Boosting 方法是一种迭代方法,在每次迭代时给不同分布赋予不同权值:增加误分类样本的权值,减少正确分类样本的权值。该方法迫使学习者对下一次迭代更多关注误分类样本。需要注意的是,Boosting 方法有效地改变了样本分布,达到了样本平衡的目的,因此被认为是一种高级重采样技术。

Weiss 提出了一种启发式、预算敏感(budget-sensitive)、渐进采样的技术来选择接近于最优的训练样本[25,26]。这种预算敏感的采样策略需要两个前提条件:一是每一类中的可用样本数量足够多,以此可用边缘分布来形成拥有 n 个样本的训练集;二是相对于获取样本时间,算法执行时间可忽略不计,而这能够保证多次执行算法直到找到所需样本。

Han 等基于预分类器提出了一种改进的过采样技术(OSPC)[27]。首先,通过训练数据获取预分类器,并保存尽可能多的多数类样本的有用信息;为提高少数类样本的分类性能,需要对少数类进行重分类。结果证明该方法要优于欠采样和 SMOTE 技术。

3 基于算法层次的解决方案

通过以上分析,重采样技术虽然简单易行,但也存在着一些难以克服的缺点。改变类分布并非是解决类不平衡问题的唯一途径。因此,企图从算法层次解决不平衡类数据的问题就愈日益迫切(即通过改变已存在的算法和技术来改变不平衡数据的特性。)

许多基于算法的解决方案力图基于内在的分类过程偏置(biasing)来解决非平衡类问题。常见的方案有:

(1)给不同类的样本赋予不同的权值[28]。

(2)通过偏好某些属性关联而对分类器偏置[29]。

(3)利用一些计数-样本对识别过程做偏置[30]。

(4)利用支持向量机的方法也设置算法偏置,从而使得超平面远离正类(稀缺类)[17]。其基本目的是补偿与不平衡数据集的稀缺关联,达到使超平面愈靠近正类数据。

在此类方案中,最典型的是代价敏感(cost-sensitive)学习、一分类分类器及集成式分类器。

3.1 代价敏感学习

传统的学习模型简单认为所有类具有相等的误分类代价。然而,在一些领域中一些特定错误的代价要远远大于其它错误。代价敏感学习方法的主要思想是将代价糅合到决策制定过程中,给与不同类赋予固定的或者是不等的误分类代价。通常误分类代价可以用代价矩阵描述,如表 1 所列。

表 1 代价矩阵

	actual negative	actual positive
predict negative	c(0,0)=c00	c(0,1)=c01
predict positive	c(1,0)=c10	c(1,1)=c11

尽管代价矩阵并不仅限于两类的情形,但为使问题更容易说明,这里用两分类问题的代价矩阵作为例子。在代价矩阵中,行代表不同的预测类别,而列代表实际的类别。矩阵中的元素 $C(i,j)$ 为把 j 类样本误分为 i 类样本的代价,$i,j\in(0,1)$。通常,错误分类正例样本(阳性)的代价为 c10,错误分类反列样本(阴性)的代价为 c01,理论上错误地给一个类标示成其它的类的代价要大于正确分类的情形,也就是 c01>c00,c10>c11。

给定代价矩阵,样本属于每个类别的条件概率为 $P(y_i|x)$,则样本应被划分到代价最小的类别,定义代价函数为 $R(y|x)$,满足下式:

$$R(y|x)=\arg\min\sum_{j=1}^{K}P(y_i|x)C(y,y_j) \tag{2}$$

在现实世界中,非同一代价是普遍存在的,作为数据挖掘领域的重大问题之一,代价敏感算法也引起了广泛关注。Maloof 认为,尽管不平衡类问题与误分类代价不同时的学习完全相同,但可以用类似途径处理[31]。

按此思路,一些工作对正类和负类的不同分类错误赋予不同的代价[32,33]。Japkowicz 等提出应用在数据集中呈现的不平衡类的比例来定义非统一错误代价[17]。Liu 等用每一类中的样本数量来规范化错误代价[34]。

3.2 一分类(one-class)分类器

与传统的分类问题(两分类或多分类)不同,一分类问题力图描述目标中的一类,且将它与其它可能外在类区分开。一分类分类器的主要思想是充分利用类不平衡问题。在此情况下,少数类可看作目标类,而多数类看作外类。目标类被较好采样意味着训练数据能较好地被映射到目标类覆盖的特征空间;与此对应,外在类能被稀疏采样,或者基本缺失。这些

类可能难以衡量，或者对此类目标衡量时代价昂贵。原则上，基于目标样本的一分类分类器可以实现。

Raskutti 等证实一分类学习在处理具有高维噪音特征空间的极不平衡数据集时显示出重要作用[35]。Juszczak 等把一分类分类器与重采样技术相结合，旨在把一些有用信息加入训练集(既包括目标类-少数类，又包括外类-多数类)[36]。

3.3 集成式分类器

集成式分类器是一种前景看好的技术，它可以改进弱分类器的性能[37]。集成式分类器通常由一组独立训练的分类器组成。集成式分类器已被证明是相当成熟的解决不平衡类问题的技术，主要因为多个独立分类器的集成的预测精度要优于单个最好的分类器预测精度。在处理不平衡类问题时，集合式分类器主要用来组合各个分类器的结果，此结果是不同分类器利用不同的过(欠)采样率进行过(欠)采样归纳得到的[38]。

3.3.1 Boosting 技术

Boosting[39,40]是一种有效的集成分类器方法，由 Freund 和 Schapire 于 1995 年提出。主要思想是从训练样例学习多个分类器，每个分类器根据前一个分类器错误分类的实例，对训练例的权重进行修正，再学习新的分类器。稳定性是 Boosting 能否发挥作用的关键因素，Boosting 可以提高不稳定学习算法的分类正确率[41]。

但是，Boosting 算法在解决实际问题时有重大的缺陷，即要求事先知道弱学习算法学习正确率的下限，这在实际问题中很难做到。1995 年，Freund 进而提出了 AdaBoost 算法，该算法可以非常容易地应用到实际问题中[42]。AdaBoost 算法的基本思想是：开始时各类数据权值相同，但是在每一轮都会增大误分类样本的权值，这样便可以使弱学习器更多地关注训练样本的少数类。正如在代价敏感数据学习中那样，通过使 AdaBoost 增权代价敏感，生成的新系统 Adacost 较 AdaBoost 产生更小的积累误分类代价，因此被广泛用于处理类不平衡问题[43]。

SMOTEBoost[44]是另一种通过改进 Boosting 技术来处理类不平衡问题的方法。该方法充分认识到过采样技术可能不利于 Boosting(如产生过适应)，因此与传统的 Boosting 通过改变每一个样本的权重以改变训练数据的分布的策略不同，SMOTEBoost 利用 SMOTE 算法通过增加新的少数类样本来改变样本分布。

3.3.2 Bagging 技术

Bagging 是一种基于组合(ensemble-based)的元学习算法，用来对数据集样本的子集进行采样，以此建立多案例学习器，并利用组合的所有预测结果决定最后类别[45]，Bagging 可以通过减少变量的平均方差来提高预测性能。

在不平衡数据集中，相对于基于 Boosting 的算法，Bagging 由于除了改变背包大小和自己采样策略外，在处理不平衡数据集时留有较少空间，因而很少引起关注。Tao 等提出了平衡采样模式只对少数类进行基采样，直到其数量与多数类等同[46]。然而，原始的 Bagging 技术选择每一个类标签独立进行基采样，结果是每个子集分布与原来的初始分布并不同。因此，基于严格平衡子集的组合模型是否能保持原类的优点尚难以确定。

Shohei 提出了大约平衡的背包技术(Roughly Balanced Bagging)来解决上述问题[47]。其主要思想在于如何决定两类中样本采样的数目：多数类样本采样的数目等于原子集样本的数目，而少数类样本采样数据取决于其二项分布。因此，同最初的背包采样平衡技术相比，利用大约平衡的背包技术采样子集的类分布就不十分平衡了。该技术的优点在于它既保留了初始背包技术的优点，同时还使得少数类样本的有用信息得到充分利用，因此较一般的背包技术鲁棒性更强。以基准数据集和真实数据集为例，该方法要超过常见的算法，如 AdaBoost，AdaBoost 与 Bagging 没有任何区别，它们最大区别在于"采样策略"[48]。

MetaCost[49]也是一种基于背包技术的合成方法。首先，它开始学习一种迭代代价敏感模型；然后，借助于 Bagging 技术估计类概率，把训练样本标记成最小期望代价类的标签；最后，利用修正后的训练集重新学习模型。

结束语 本文讨论了数据集不平衡分类问题产生的原因、已有的几类主要解决方案。由于数据集不平衡问题是随着数据获取、机器学习和信息检索的发展而出现的一个比较新的问题，对该问题的研究并不成熟，未来的研究热点聚集在以下几个方面：

(1)当前，大多数针对不平衡问题主要着眼于决策树分类器，而相关研究结果证明其并不适用于神经网络、回归、最近邻居等，如何产生适用于这些分类器的算法尚待研究。

(2)从本质上讲，不平衡分类问题并非是由于数据集不平衡直接导致的，相反，类不平衡性会产生一些小数据块，而反过来加剧不平衡性。尽管在处理不平衡类内部和不平衡类之间设置的最大偏置对解决这类问题产生了一定效用[50]，但如何生成更有效的方法是下一步研究的重点。

(3)针对不平衡分类问题已经出现了若干解决方案和思路，但其都是针对类间不平衡问题的。在许多情况下，除了类间不平衡问题外，类内样本的不平衡问题同样影响着分类精度，不平衡类问题的解决既要着眼于类间不平衡也要关注类内不平衡问题对分类性能的影响。如何应用有效的知识指导类内不平衡问题尚待探讨。

(4)迄今为止，不平衡分类问题还局限在算法层次上，缺乏相关理论基础，研究成果还很少，多数研究也是依据实验方法，实验结果也多是经验性的，缺乏成熟理论的支撑。因此进一步的理论研究非常重要。

参 考 文 献

[1] Chan P K, Stolfo S J. Toward scalable learning with nonuniform class and cost distributions: A case study in credit card fraud detection[C] // Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York, USA: AAAI Press, 1998: 164-168

[2] Phua C, Alahakoon D, Lee V. Minority report in fraud detection: Classification of skewed data[J]. SIGKDD Explore, 2004, 6(1): 50-59

[3] Sun Aixin, Lim E P, Liu Ying. On strategies for imbalanced text classification using SVM: A comparative study[J]. Decision Support Systems, 2009, 48: 191-201

[4] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000, 2(4): 303-336

[5] Ling C X, Li C. Data mining for direct marketing: Problems and

solutions[C]// Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining. 1998:73-79

[6] Bauer E, Kohavi R. An empirical comparison of voting classication algoirthm: Bagging, boosting and variants[J]. Machine Learning, 1999, 36:105-142

[7] Japkowicz N, Stephen S. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis Journal, 2002, 6(5): 429-450

[8] Joshi M V. Learning Classier Models for Predicting Rare Phonemena[D]. University of Minnesota USA, 2002

[9] Chawla N V, Japkowicz N, 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

[10] Weiss G, 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

[11] Chawla N V, Hall L O, Bowyer K W, et al. SMOTE: Synthetic Minority Oversampling TEchnique[J]. Journal of Artificial Intelligence Research, 2002, 16:321-357

[12] Batista G E, 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

[13] Han Hui, Wang Wenyuan, Mao Binghuan. Borderline-smote: a new oversampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science, 2005, 3644:878-887

[14] Liu Alexander, Ghosh J, Martin C. Generative oversampling for mining imbalanced datasets[C]// Proceedings of 2007 International Conference on Data Mining. Las Vegas, Nevada, USA: CSREA Press, 2007:66-72

[15] Batista G E, Pratti R C, Monard C A. Study of the behavior of several methods for balancing machine learning training data [J]. SIGKDD Explorations, 2004, 6:20-29

[16] Barandela R, Hernandez J K, Sanchez J S, et al. Imbalanced training set reduction and feature selection through genetic optimization[C]// Proceeding of the 2005 Conference on Artificial Intelligence Research and Developments. 2005:215-222

[17] Japkowicz N. Concept-learning in the presence of between-class and withinclass imbalances[C]// Proceedings of 14th Conference of the Canadian Society for Computational Studies of Intelligence. Ottawa, Canada, 2001:67-77

[18] Zhang Jianping, Mani I. knn approach to unbalanced data distributions: A case study involving information extraction[C] // Machine Learning Workshop on Learning from Imbalanced Datasets. Washington DC, USA, 2003

[19] Kubat M, Matwin S. Addressing the Curse of Imbalanced Data Sets: One Sided Sampling[C]// Proceedings of the Fourteenth International Conference on Machine Learning. San Fransisco, USA: 1997:179-186

[20] Batista G, Carvalho A, Monard M C. Applying Onesided Selection to Unbalanced Datasets[C]// Proceedings of the Mexican International Conference on Artificial Intelligence. Acapulco, Mexico: Springer-Verlag, 2000:315-325

[21] Barandela R, Valdovinos R M, Sanchez J S, et al. The imbalanced training sample problem: Under or over sampling? [C]// Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR'04), Lecture Notes in Computer3138. 2004:806-814

[22] Barandela R, Rangel E, Sanchez J S, et al. Restricted decontamination for the imbalanced training sample problem[C]// Proceedings of the 8th Ibero-american Congress on Pattern Recognition. Havana, Cuba, 2003:424-431

[23] Ling C X, Li Chenghui. Data Mining for Direct Marketing Problems and Solutions[C]// Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York, USA: AAAI Press, 1998:73-79

[24] Zheng Zhaohui. Feature selection for text categorization on imbalanced data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1):80-89

[25] Weiss G M. The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning[D]. Rutgers University, USA, 2003

[26] 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

[27] 韩慧, 王路, 温明, 等. 均衡数据集学习中基于初分类的过抽样算法[J]. 计算机应用, 2006, 26(8):1894-1897

[28] Pazzani M, Merz C, Murphy P, et al. Reducing misclassification costs[C]// Proceedings of 11th International Conference on Machine Learning. New York, USA, 1994:217-225

[29] Ezawa K J, Singh M, Norton S W. Learning goal oriented Bayesian networks for telecommunications management[C]// Proceedings of 13th International Conference on Machine Learning. Bari, Italy, 1996:139-147

[30] Kubat M, Holte R, Matwin S. Detection of oil-spills in radar images of sea surface[J]. Machine Learning, 1998, 30:195-215

[31] Maloof M A. Learning when data sets are imbalanced and when costs are unequal and unknow[C]// Workshop on Learning from Imbalanced Data Sets. Menlo Park, CA: AAAI Press, 2003

[32] Domingos P. Metacost: a general method for making classifiers cost-sensitive[C]// Proceedings of 5th International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM Press, 1999:155-164

[33] Gordon D F, Perlis D. Explicitly biased generalization[J]. Computational Intelligence, 1989, 5:67-81

[34] Yan Rong, Liu Yan, Rong Jin, et al. On predicting rare classes with SVM ensembles in scene classification[C]// Proceedings of International Conference on Acoustics, Speech and Signal Processing. Hongkong: IEEE Press, 2003:21-24

[35] Raskutti B, Kowalczyk A. Extreme rebalancing for svms: a case study[J]. SIGKDD Explorations, 2004, 6:60-69

[36] Juszczak P, Duin P W. Uncertainty sampling methods for one-class classifiers[C]// Proceedings of International Conference on Machine Learning, Workshop on Learning with Imbalanced Data Sets II. 2003:81-88

[37] 张琦, 吴斌, 王柏. 非平衡数据训练方法概述[J]. 计算机科学, 2005, 32(10):181-186

[38] Kotsiantis S, Pintelas P. Mixture of expert agents for handling imbalanced data sets. Annals of Mathematics[J]. Computing & TeleInformatics, 2003, 1:46-55

[39] Schapire R E, Freund Y, Bartlett Y, et al. Boosting the margin:

A new explanation for the effectiveness of voting methods[C]// Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1997:322-330

[40] Freimd Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121(2):256-285

[41] Quinlan J R. Bagging, Boosting and C4 [C]//Proceedings of the 13th International Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1996:725-730

[42] 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

[43] Fan Wei, Stolfo S J, Zhang Junxin, et al. AdaCost: misclassification cost-sensitive boosting[C]// Proceedings of the Sixteenth International Conference on Machine Learning. San Mateo, USA, 1999:99-105

[44] Chawla N V, Lazarevic A, Hall L O, et al. Smoteboost: Improving prediction of the minority class in boosting [C]// Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Database. Dubrovnik, Croatia, 2003:107-119

[45] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140

[46] 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): 1088-1099

[47] Hido S, Kashima H. Roughly Balanced Bagging for Imbalanced Data[C]// Proceedings of the SIAM International Conference on Data Mining. Atlanta, USA, 2008:143-152

[48] 毕华,梁洪力,王珏. 重采样方法与机器学习[J]. 计算机学报, 2009,32(5):862-877

[49] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]// Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM Press, 1999:155-164

[50] Zheng Zhaohui, Wu Xiaoyun, Srihar R. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004, 6(1):80-89

(上接第 22 页)

[66] Bertoni A, Folgieri R, Valentini G. Bio-molecular cancer prediction with random subspace ensembles of support vector machines[J]. Neurocomputing, 2005, 63(1):535-539

[67] Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006, 7:3

[68] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification[J]. BMC Bioinformatics, 2008, 9:319

[69] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1): 239-263

[70] Peng Y H. A novel ensemble machine learning for robust microarray data classification[J]. Computers in Biology and Medicine, 2006, 36(6):553-573

[71] Kim K J, Cho S B. An Evolutionary Algorithm Approach to Optimal Ensemble Classifiers for DNA Microarray Data Analysis [J]. IEEE Trans on Evolutionary Computation, 2008, 12(3): 377-388

[72] Yeang C H, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types[J]. Bioinformatics, 2001, 17(1): 316-322

[73] Lee Y, Lee C. Classification of multiple cancer types by multicategory support vector machines using gene expression data[J]. Bioinformatics, 2003, 19(9):1132-1139

[74] Berrar D, Bradbury I, Dubitzky W. Instance-based concept learning from multiclass DNA microarray data[J]. BMC Bioinformatics, 2006, 7:73

[75] Shen L, Tan E C. Reducing multiclass cancer classification to binary by output coding and SVM[J]. Computational Biology and Chemistry, 2005, 30(1):63-71

[76] Statnikov A, Aliferis C F, Tsamardinos I, et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis[J]. Bioinformatics, 2005, 21(5):631-643

[77] Braga-Neto U M, Dougherty E R. Is cross-validation valid for small-sample microarray classification? [J]. Bioinformatics, 2004, 20(3):374-380

[78] Fu W J, Carroll R J, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation[J]. Bioinformatics, 2005, 21(9):1979-1986

[79] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection[J]. BMC Bioinformatics, 2006, 7:91

[80] Wood I A, Visscher P M, Mengersen K L. Classification based upon gene expression data: bias and precision of error rates[J]. Bioinformatics, 2007, 23(11):1363-1370

[81] Dougherty E R, Hua J P, Bittner M L. Validation of Computational Methods in Genomics[J]. Current Genomics, 2007, 8(1): 1-19

[82] Hua J P, Xiong Z X, Lowey J, et al. Optimal number of features as a function of sample size for various classification rules[J]. Bioinformatics, 2005, 21(8):1509-1515

[83] Jiang H Y, Deng Y P, Chen H S, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes[J]. BMC Bioinformatics, 2004, 5:81

[84] Yoon Y, Lee J, Park S, et al. Direct integration of microarrays for selecting informative gene and phenotype classification[J]. Information science, 2008, 178(1/2):88-105

[85] Vogiatzis D, Tsapatsoulis N. Active learning for microarray data [J]. International Journal of Approximate Reasoning, 2008, 47(1):85-96