第 42 卷 第 2 期
2006 年 3 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCES)

Vol. 42, No. 2
Mar., 2006

一种基于级联模型的类别不平衡数据分类方法*

刘胥影[1], 吴建鑫[2], 周志华[1]**
(1. 南京大学软件新技术国家重点实验室, 南京, 210093;
2. 佐治亚理工学院计算机学院, 美国佐治亚州亚特兰大, 30332-0280)

摘 要: 真实世界问题中,不同类别的样本在数目上往往差别很大,而传统机器学习方法难以对小类样本进行正确分类,若小类的样本是足够重要的,就会带来较大的损失.因此,对类别分布不平衡数据的学习已成为机器学习目前面临的一个挑战.受计算机视觉中级联模型的启发,提出一种针对不平衡数据的分类方法 BalanceCascade.该方法逐步缩小大类别使数据集趋于平衡,在此过程中训练得到的一系列分类器通过集成方式对预测样本进行分类.实验结果表明,该方法可以有效地提高在不平衡数据上的分类性能,尤其是在分类性能受数据的不平衡性严重影响的情况下.

关键词: 机器学习, 数据挖掘, 类别不平衡, 级 联, 集成学习
中图分类号: TP 18

A Cascade-based Classification Method for Class-imbalanced Data

Liu Xu-Ying[1], *Wu Jian-Xin*[2], *Zhou Zhi-Hua*[1]
(1. National Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, China;
2. College of Computing, Georgia Institute of Technology, Atlanta, 30332-0280, USA)

Abstract: In machine learning and data mining, there are many aspects that might influence the performance of a learning system in real world applications. Class imbalance is one of these factors, in which training examples in one class heavily outnumber the examples in another class. Classifiers generally have difficulty in learning concept from the minority class. In many applications if the minority class is more important than the majority class, there will be great loss.

There is severe class imbalance in the face detection problem, which greatly decreases the detection speed. The cascade structure is proposed to accelerate the learning process. Cascade is a classifier system with a sequence of n node classifiers. At the beginning, all training examples are available to train the first node classifier. Then all positive examples and only a subset of negative examples are passed to the next node, neglecting those negative examples correctly classified by the first node. This procedure repeats until all node classifiers are trained. A test example is passed to the next node if it is recognized as positive by the current node, or is rejected immediately as negative. However, the learning goal of a cascade node classifier is quite different to usual classifiers in the sense that every node aims to get a high detection rate and only a moderate false alarm rate. The cascade can achieve both high overall detection rate and low overall false alarm rate.

* 基金项目:国家杰出青年科学基金(60325207),江苏省自然科学基金重点项目(BK2004001),
"973"国家计划(2002CB312002)
收稿日期:2005-09-04
** 通讯联系人,E-mail:zhouzh@nju.edu.cn

Every time training examples are passed to the next node, there are some negatives that are neglected. That is, there are fewer negatives in the training set than those in the previous node. Considering the class imbalance problem, it means a more balanced training set, compared with training sets in previous nodes. In early nodes within a cascade it is quite easy to achieve the learning goal, i. e. train a classifier with high detection rate and only moderate false alarm rate. However, it becomes harder in deeper nodes, since the negative examples in these nodes are false positives from previous nodes and are difficult to separate from positive examples. And there's another difference between the face detection problem and general class imbalance problems. Hundreds of thousands of features are available for classifiers in the former case, but it is not the case for the latter one. In general class imbalance problems, a classifier in a deeper node may not easily achieve both a high detection rate and a moderate false alarm rate. Therefore, cascade-style test may not be appropriate in general class imbalance problems. Instead of testing new examples in a cascade sequential style, we combine all the node classifiers into an ensemble classifier and propose a cascade-based classification algorithm, BalanceCascade, to deal with class imbalance problems. Particularly, BalanceCascade employs Adaboost to train a classifier in each node, which is a weighted combination of several weak learners. Then weak learners within all node classifiers are collected to form the final ensemble without changing their original weights. Experimental results show that the method can effectively improve the classification performance on imbalanced data sets, especially in the cases when classification performance is heavily affected by class imbalance.

Key words: machine learning, data mining, class imbalance, cascade, ensemble learning

在机器学习和数据挖掘研究中,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真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多于其他类别.在这种情况下,分类器通常会倾向于将测试样本全部判别为大类而忽视小类,这使得到的分类器在小类上效果很差.根据不同的问题,大类和小类的样本数之比可达 100 甚至 10 000 以上[1,2],并且这种不平衡问题普遍存在于很多领域,例如金融欺诈检测[3]、石油勘探[4]、语音处理[5]、信息检索[6]、反垃圾邮件[7]、医学诊断、网络入侵检测等.例如,在网络入侵检测问题中,入侵的数量只占整个网络流量的极少一部分.这些应用常常要求小类的检测率足够高,同时要求大类的错误率足够小.

分类器在不平衡数据上性能下降的原因有很多[8],例如不恰当的性能评价准则、不恰当的归纳偏置、由于一类样本数目过少产生的绝对稀少(rarity)问题、由于各类样本数目相差悬殊产生的相对稀少问题、以及采取分而治之(divide and conquer)策略算法所独有的数据碎片(data fragmentation)问题和噪音等.现有的对策大致包括[8]选择合适的性能评价准则、采用非贪婪的搜索策略、选择合适的归纳偏置、与专家或知识交互、分割数据以降低数据的不平衡性、通过取样方法改变数据的原始分布、只对一类进行学习、利用代价敏感学习(cost sensitive learning)解决不平衡问题等.

在计算机视觉中,人脸检测是一个重要的研究问题,一副图像中包含人脸的图像区域非常少.为了提高检测速度,Viola 和 Jones[9]提出了 AdaBoost 级联模型并获得了成功.受该思想的启发,本文提出了一种新的针对类别不平衡数据的学习方法 BalanceCascade.该方法逐步缩小大类使数据趋于平衡,在此过程中训练得到的一系列分类器集成起来对测试样本进行分类.实验结果表明,该方法可以有效提高分类器在不平衡数据上的性能.

1　AdaBoost 算法

集成学习利用多个学习器解决问题可以获得较强的泛化能力,目前该技术已被成功应用到肺癌细胞识别等领域[10].AdaBoost[11]是一种著名的集成学习算法,可以有效提高单一学

习器的泛化能力. 其中每个训练样本都有初始权值,在每轮训练一个弱分类器后分类正确的样本权值变小,分类错误的样本权值变大. 最后这些弱分类器集成为一个强分类器完成分类任务.

何谓弱分类器? 在 PAC 学习理论[12]中, PAC - 强可学习算法是指:给定 $\varepsilon,\delta>0$,在合理时间内可以以 $1-\delta$ 的概率得到错误率小于 ε 的假设;限制 $\varepsilon\geq 1/2-\gamma$ 便可得到 PAC - 弱可学习算法,其中 γ 为大于 0 的常数.

若 X 是问题的特征空间, x_i 为表示第 i 个样本的特征向量, y_i 为该样本的类别标记,其值为 0 或 1,则有 N 个训练样本的数据集为 $\{x_i, y_i\}$,其中 x_i 是从 X 中随机取样得到的. 分类器的任务是从训练集中学到一个假设 $h: X\to\{0,1\}$. 也可以输出概率形式 $h: X\to[0,1]$, h 值的大小表示分类结果为 1 的置信度. 衡量一个分类器的性能可用错误总数的期望值: $E_{x\sim D}(|h(x)-y|)$. 具体算法描述请参见[11].

2 BalanceCascade 方法

受到 Viola 和 Jones[9]以及 Wu 等人工作[13]的启发,本文提出了基于级联模型(Cascade)的 BalanceCascade 方法来解决数据的不平衡问题.

一副图像中包含了几百万个要检测的图像区域,但其中只有很少一些对应人脸. 若对所有这些可能的区域采用同样复杂的特征测试,势必会大大降低检测速度. 为此, Viola 和 Jones 提出了级联模型[9]. 如图 1 所示,这是由一组分类器 $\{H_i\}_{i=1}^{n}$ 构成的层次结构. 每个结点上的分类器 H_i 要有很高的检测率(detection rate,图中表示为 d_i)和适当大小的误警率(false alarm rate,反例中被错分为正例的比例,图中表示为 f_i) ,比如 0.5. 输入的图像区域若被检测为包含人脸图像,则由 H_i 传递给下一级的 H_{i+1} 继续检测,否则就被拒绝. 最终的检测率 d 和误警率 f 分别为 $\prod_{i=1}^{n} d_i$ 和 $\prod_{i=1}^{n} f_i$. 分类器是逐次得到的,从 H_1 开始. 其后每个结点获得由上一个结点传递的所有的正类样本和反类的一个子集作为该结点的训练集. 其中,反类样本中被正确分类时不再传递到下一结点继续训练分类器. 测试样本被分类为正类时由下一个结点的分类器继续测试,否则为反类.

每个结点的分类器只要求有足够高的检测率和适当大小的错误正类率,这一点是相对容易满足的. 因此,在前面的结点中只使用少量的特征就可以得到一个满足检测率和误警率的要求的分类器. 越往后,虽然要使用的特征越多,但同时反类样本也越来越少. 因此级联模型可以有效的提高检测性能. 在 Viola 和 Jones 的工作之后, Wu 等人[13]也对级联模型进行了研究,并得到了效率更高的人脸检测方法.

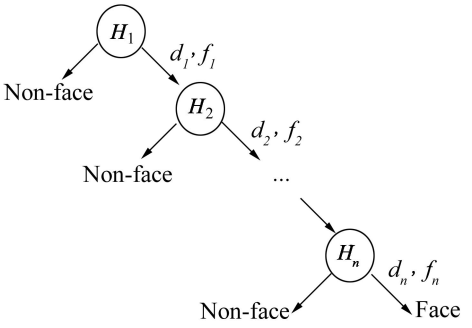

图 1 有 n 个结点的级联模型示意图

Fig 1 Illustration of cascade model with n nodes

在级联结构中,每次训练样本从一个结点传到下一个结点时总有一些反类样本被扔掉. 也就是说,每个结点训练集中的反类样本数比上一个结点减少了. 当级联结构面临的是不平衡数据时,这意味着得到了更为平衡的训练集. 级联结构结点中数据的不平衡程度越大,就有越多的反类样本被拒绝在外. 所以越是前面的结点,越容易使下一结点的不平衡程度减弱. 同时,随着越来越多的易于分类的反类样本被排除,分类任务变得越来越困难. 这一点对于人脸检测问题同样成立. 在人脸检测问题中,由于可以得到成千上万的样本特征,所以即使当分类任务困难时也总是能够得到性能较好的分类器,从而保证了总体有很高的检测率和很小的

误警率.但是在数据不平衡问题中不具备这样的优势,当分类任务困难时,很难得到一个同时有很高检测率和中等程度误警率的分类器,因此分层测试的方式不再适用,我们可以将级联结构中得到的所有分类器集成起来作为一个新的分类器.综合考虑以上几点,我们提出了针对不平衡数据的分类算法 BalanceCascade.并且,由于面对的是不平衡的数据集,正类样本是相对稀少的,因此只要保持误警率在合适的大小(如略高于 0.5),就可以得到很高的检测率.除此之外,我们注意到,在级联结构中的每个结点中,不平衡程度虽然减弱但依然存在,为了得到可靠的分类器,可以对该结点的训练样本应用随机下取样(random under-sampling)方法,即在其中随机选出部分反类样本使其数目和正类相同,这部分反类样本和所有正类样本组成该结点的训练集,然后调用 AdaBoost 算法训练分类器.如本文第 1 部分所述,AdaBoost 算法得到的分类器是多个弱分类器的加权组合.在 BalanceCascade 算法中,我们按以下方式将各结点对应的分类器集成起来:将其中所有的弱分类器加权结合,同时保持其原有权值不变.BalanceCascade 算法的描述具体如表 1 所示.

表 1　BalanceCascade 算法

Table 1　BalanceCascade algorithm

Algorithm BalanceCascade

给定:数据集,其中正类样本集合为 P,反类样本集合为 N,并且正类样本数远小于反类样本数:$|P| \ll |N|$

　　分类器的误警率 α

　　Cascade 结构的结点数 n

　　Adaboost 算法的轮数 m

结点数 $i = 0$

Do While $|P| < |N|$

　1. 结点数 $i = i + 1$

　2. 从 N 中随机取样得到 N',并且 $|N'| = |P|$

　3. 由 N 和 P 组成训练集,使用 AdaBoost 算法训练得到一个分类器 H_i,其中的弱分类器集合为 $\{h_{i,j}\}_{j=1}^{m}$

　4. 调整 H_i 的阈值使误警率为 α

　5. 用分类器 H_i 对反类样本集合 N 进行分类,并从 N 中移除被正确分类的样本

　保持权值不变,将所有结点分类器 $\{H_i\}_{i=1}^{n}$ 内的弱分类器 $\{h_{i,,j}\}_{(i,j)=(1,1)}^{(n,m)}$ 加权集成为最终的分类器 H

　输出假设

$$H(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\log\frac{1}{\beta_{i,j}}\right)h_{i,j}(x) \geq \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\log\frac{1}{\beta_{i,j}} \\ 0 & \text{otherwise} \end{cases}$$

　其中,$\beta_{i,j}$ 是弱分类器 $h_{i,j}$ 在 AdaBoost 算法中对应的 β 值

3　实　验

3.1　比较方法　AdaBoost 方法作为比较基准,除此之外,取样方法(sampling)是解决数据不平衡问题的常用方法[14].一种是简单常用的随机下取样方法(简记为 UnderSampl),该方法随机取出部分反类样本与所有正类组成新的训练集,其中取出的反类样本数目与正类相同.与下取样方法相对应的是上取样方法(over-sampling),它通过增加小类样本的方式使数据达到平衡.SMOTE[15] 是一种常用的上取样方法,它通过在小类样本和其同类近邻间插值生成人工样本的方式扩大小类.我们实现一种混合的取样方法 HSampl(Hybird Sampling),先

用 SMOTE 方法将小类样本增加一倍,然后用随机下取样缩小大类使数据集平衡. UnderSampl 和 HSampl 方法对训练集操作后,用 AdaBoost 算法训练分类器.

AdaBoost 对正类和反类的分类错误同等对待,这是导致其在不平衡数据上性能下降的一个重要原因. 为此,Viola 和 Jones 提出了 AdaBoost 的一个变体 Asymboost[16],在此算法中,正类样本被错误分类时代价更高. 当正类和负类具有相同的错误分类代价时,该算法退化为 AdaBoost 算法.

3.2 实验设置 本文在 10 个 UCI[17] 数据集上进行测试,有关信息见表 2. 其中有一些是多类的数据集,本文选择其中的一类为正类,剩余的为反类. 实验方法为 5 次 10 倍交叉验证.

表 2 实验中使用的数据集

Table 2 Experimental datasets

Dataset	Size	Class	Positive class	|P| ∶|N|
Abalone	4 177	29	Ring = 7	391 ∶3 786
Car	1 728	4	acc	384 ∶1 344
Cmc	1 473	3	Class = 2	333 ∶1 140
Haberman	306	2	Class = 2	81 ∶225
Ionosphere	351	2	good	126 ∶225
Letter	20 000	26	A	789 ∶19 211
Phoneme	5 427	2	Class = 1	1 586 ∶3 818
Pima	768	2	Class = 1	268 ∶500
Satimage	6 435	6	Class = 4	626 ∶5 809
Wdbc	569	2	Class = *M*	212 ∶357

Adaboost 的弱分类器采用 Breiman 等人提出的决策树[18]方法. Asymboost 中正类和反类的错分代价之比等于反类和正类的样本数之比. AdaBoost、UnderSampl、HSampl 和 Asymboost 方法均训练 40 轮,BalanceCascade 方法使用的预设误警率见表 3,每个结点的分类器训练 10 轮. 如此 BalanceCascade 和其他方法中所含的弱分类器的数目大致相当,可以保证比较的相对公平性.

表 3 BalanceCascade 算法中预设的误警率

Table 3 Values of false positive rates in BalanceCascade

	误警率		误警率
Abalone	0.5	Letter	0.4
Car	0.7	Phoneme	0.75
Cmc	0.7	Pima	0.8
Haberman	0.7	Satimage	0.5
Ionosphere	0.8	Wdbc	0.7

本文采用查准率、查全率和 F 值[19,20]这几个准则对分类器的性能进行评估.

3.3 实验结果及分析 实验结果如表 4 所示,在所有数据集上的平均结果如表 5 所示,其中每个数据集上最好的结果用粗体标出,最差的结果以“ * ”标出.

在表 4 中可以看出,数据的不平衡性对分类器性能的影响不是绝对的,如 letter 数据集中反类和正类样本数目之比接近 25 ∶1,但 Adaboost 算法的查准率、查全率和 F 值都很高. 除此之外,在 car、ionosphere 和 wdbc 上,Adaboost 算法的分类性能也很高,其 F 值都在 0.9 以上.

Adaboost 算法更关注于查全率 Precision,具体来说,它在 6 个数据集上的查全率最高,平均的查全率在所有的方法中是也是最高的. 但其查全率很低,在 8 个数据集上其查全率最低,结果平均查全率在所有方法中最差. 由于过低的查全率,所以即使查准率很高,F 值仍然很

低,仅为 0.690. UnderSampl 算法正好与 Adaboost 算法相反,它倾向于获得更高的查全率而忽视了查准率. 它在 8 个数据集上查全率最高,而查准率在 9 个数据集上最低,对应的 F 值甚至不如对不平衡数据采取任何策略的 Adaboost 方法.

综合以上分析,在不平衡的 10 个数据集上,分类器按以 F 值为准则的分类性能由高到低排列依次为: BalanceCascade、HSampl、Asymboost、Adaboost 和 UnderSampl. 其中 UnderSampl 算法和 Asymboost 算法并没有有效地改善 Adaboost 算法在不平衡数据上的分类性能. UnderSampl 算法性能变差的原因可能在于由于大大缩小了反类而导致可用信息减少,由此引起分类性能的下降. Asymboost 算法中,错误分类正类样本和反类样本的代价不同,这说明利用代价信息解决数据的不平衡问题未必是有效的. HSampl 算法实际上结合了上取样和下取样,一方面通过增加样本的方式强调了正类,一方面对反类进行适当程度的缩小,可以有效的提高分类器的性能. BalanceCascade 方法的平均分类性能最好,尤其可以显著提高严重受数据不平衡性影响的学习算法的分类能力. BalanceCascade 方法中有 3 个要素,一是根据利用级联结构逐步缩小反类使数据集趋向平衡,二是在各个结点用下取样方法对训练集进行操作,三是集成机制. 由 UnderSampl 方法可以推理得到,第二个要素未必对 BalanceCascade 方法有所贡献,真正起作用的是利用级联结构逐级使数据集趋向平衡的策略和集成机制.

表 4　10 个 UCI 数据集上的 precision、recall 和 F 值

Table 4　Precision, recall and F-value on 10 UCI data sets

(Abalone)	Precision	Recall	F-value	(Car)	Precision	Recall	F-value
AdaBoost	0.288	0.169 *	0.210 *		**0.962**	0.972 *	**0.967**
Asymboost	**0.291**	0.182	0.222		0.960	0.974	0.966
UnderSampl	0.239 *	**0.795**	0.367		0.805 *	0.984	0.884 *
HSampl	0.261	0.693	0.378		0.893	0.974	0.931
BalanceCascade	0.261	0.728	**0.384**		0.858	**0.987**	0.917
(Cmc)	Precision	Recall	F-value	(Haberman)	Precision	Recall	F-value
AdaBoost	**0.397**	0.385 *	0.388 *		0.347	0.363 *	0.354 *
Asymboost	0.386	0.420	0.400		0.343 *	0.391	0.365
UnderSampl	0.328 *	**0.625**	0.429		0.355	**0.601**	**0.442**
HSampl	0.371	0.479	0.417		0.362	0.473	0.406
BalanceCascade	0.349	0.585	**0.436**		**0.363**	0.565	0.439
(Ionosphere)	Precision	Recall	F-value	(Letter)	Precision	Recall	F-value
AdaBoost	0.949	0.875 *	0.907		**0.998**	0.978	**0.988**
Asymboost	**0.953**	0.876	**0.910**		**0.998**	0.977 *	0.987
UnderSampl	0.919 *	**0.888**	0.900 *		0.827 *	**0.997**	0.903 *
HSampl	0.935	0.885	0.907		0.918	0.994	0.954
BalanceCascade	0.932	0.886	0.905		0.960	0.994	0.976
(Phoneme)	Precision	Recall	F-value	(Pima)	Precision	Recall	F-value
AdaBoost	**0.862**	0.839 *	0.850		**0.631**	0.599 *	0.611 *
Asymboost	0.860	0.845	**0.852**		0.629	0.605	0.613
UnderSampl	0.748 *	**0.906**	0.819 *		0.579 *	**0.730**	0.644
HSampl	0.815	0.883	0.847		0.615	0.647	0.627
BalanceCascade	0.793	0.892	0.839		0.603	0.708	**0.649**
(Satimage)	Precision	Recall	F-value	(Wdbc)	Precision	Recall	F-value
AdaBoost	**0.780**	0.580 *	0.664		0.969	0.946	0.956
Asymboost	0.773	0.592	**0.668**		**0.970**	0.945 *	0.956
UnderSampl	0.395 *	**0.890**	0.546 *		0.945 *	0.961	0.952 *
HSampl	0.486	0.820	0.610		0.966	0.951	**0.957**
BalanceCascade	0.495	0.829	0.619		0.953	**0.964**	**0.957**

黑体为每个数据集上最好的结果,“ * ”为最差的结果

表 5 平均结果

Table 5 Averaged results

average	Precision	Recall	*F*-value
AdaBoost	**0.718**	0.671 *	0.690
Asymboost	0.716	0.681	0.694
UnderSampl	0.614 *	**0.838**	0.689 *
HSampl	0.662	0.780	0.703
BalanceCascade	0.657	0.814	**0.712**

黑体为每个数据集上最好的结果,“ * ”为最差的结果

4 总 结

受级联结构的启发,本文提出了一种解决数据不平衡性的新型分类方法 BalanceCascade. 该方法利用级联结构逐步缩小大类使数据集趋于平衡,在此过程中得到的一系列分类器集成起来完成分类任务. 在 10 个数据集上的实验结果表明,BalanceCascade 可以有效的提高分类性能,尤其是在分类性能严重受到数据的不平衡性影响的情况下.

在本文提出的 BalanceCascade 算法中,各结点的分类器以简单的方式组合在一起成为一个强的分类器,即保持所有分类器内的弱分类器的权值不变,直接将它们加权组合. 是否存在更有效的集成方式将是今后研究的问题.

References

[1] Pearson R, Goney G, Shwaber J. Imbalanced clustering for microarray time-series. Proceedings of the ICML '03 Workshop on Learning from Imbalanced Data Sets. Washington, DC, 2003.

[2] Wu G, Chang E Y. Class-boundary alignment for imbalanced dataset learning. Proceedings of the ICML '03 Workshop on Learning from Imbalanced Data Sets. Washington, DC, 2003.

[3] Chan P K, Stolfo S J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. NY: AAAI Press, 1998: 164～168.

[4] Kubat M, Holte R C, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 1998, 30(2): 195～215.

[5] Van den Bosch A, Weijters T, van den Herik H J, *et al*. When small disjuncts abound, try lazy learning: A case study. Proceedings of the 7th Belgian-Dutch Conference on Machine Learning. Tilburg: Tilburg University Press, 1997, 109～118.

[6] Lewis D, Catlett J. Heterogeneous uncertainty sampling for supervised learning. Proceedings for the 11th International Conference of Machine Learning. New Brunswick, NJ: Morgan Kaufmann Press, 1994, 148～156.

[7] Fawcett T. “In vivo ”spam filtering: A challenge problem for data mining. SIGKDD Explorations, 2003, 5(2): 140～148.

[8] Weiss G. Mining with rarity: A unifying framework. SIGKDD Explorations, 2004, 6(1): 7～19.

[9] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA, 2001: 511～518.

[10] Jiang Y, Zhou Z H, Xie Q, *et al*. Application of neural network ensemble in lung cance cell identification space. Journal of Nanjing University (Natural Sciences), 2001, 37(5): 529～534. (姜 远, 周志华, 谢 琪等. 神经网络集成在肺癌细胞组识别中的应用. 南京大学学报(自然科学), 2001, 37(5): 529～534).

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

[12] Kearns M J, Vazirani U V. An introduction to computational learning theory. Cambridge, MIT Press, 1994.

[13] Wu J, Rehg J M, Mullin M D. Learning a rare event detection cascade by direct feature selection. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2003.

[14] Batista G E A P A, Prati R C, Modard M C. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, 2004, 6(1): 20～29.

[15] Chawla N V, Bowyer K W, Hall L O, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321～357.

[16] Viola P, Jones M. Fast and robust classification using asymmetric AdaBoost and a detector cascade. Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002: 1 311～1 318.

[17] Blake C, Keogh E, Merz C J. UCI repository of machine learning databases. http://www.ics.uci.edu/～mlearn/MLRepository.html, Department of Information and Computer Science, University of California, Irvine, 1998.

[18] Breiman L, Friedman J H, Olshen R A, *et al*. Classification and regression trees. Wadsworth: Oxford University Press, 1984.

[19] Buckland M, Gey F. The relationship between recall and precision. Journal of the American Society for Information Science, 1994, 45(1): 12～19.

[20] Joshi M, Kumar V, Agarwal R. Evaluating boosting algorithms to classify rare classes: Comparison and Improvements. Proceedings of the 1st IEEE International Conference on Data Mining. 2001, 257～264.