计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
47(8): 1407-1414, 2010

类别不平衡的分类方法及在生物信息学中的应用

邹 权 郭茂祖 刘 扬 王 峻
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)
(zouquan@xmu.edu.cn)

A Classification Method for Class-Imbalanced Data and Its Application on Bioinformatics

Zou Quan, Guo Maozu, Liu Yang, and Wang Jun
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract A classification method is proposed for class-imbalanced data, which is common in bioinformatics, such as identifying snoRNA, classifying microRNA precursors from pseudo ones, mining SNPs from EST sequences, etc. It is based on the main idea of ensemble learning. First, the big class set is divided randomly into several subsets equally, and it is made sure that every subset together with the small class set can make up a class-balanced training set. Then several different mechanism classifiers are selected and trained with these balanced training sets. After the multi-classifiers are built, they will vote for the last prediction when dealing with new samples. In the training phase, a strategy similar to AdaBoost is used. For each classifier, the samples will be added to the training sets of next two classifiers if they are misclassified. It is necessary to repeat modifying the training sets until a classifier can accurately predict its training set or reaching the maximum repeat times. This strategy can improve the performance of weak classifiers by voting. Experiments on five UCI data sets and three bioinformatics experiments mentioned above prove the performance of the method. Furthermore, a software program named LibID, which can be used as similarly as LibSVM, is developed for the researchers from bioinformatics and other fields.

Key words bioinformatics; class imbalance; ncRNA identification; mining SNP from EST; classification

摘 要 提出一种处理正反例不平衡的分类方法,以解决生物信息学中的 snoRNA 识别、microRNA 前体判别、SNP 位点的真伪识别等问题. 利用集成学习的思想,将反例集均匀分割并依次与正例集组合,得到一组类别平衡的训练集. 然后对每个训练集采用不同原理的分类器进行训练,最后投票表决待测样本. 为了避免弱分类器影响投票效果,结合 AdaBoost 思想,将每个分类器训练中产生的错误样本加入到下 2 个分类器的训练集中,既避免了 AdaBoost 的反复训练,又有效地利用投票机制遏制了弱分类器的影响. 5 组 UCI 测试数据和 3 组生物信息学实验证明了它在处理类别不平衡分类问题时的优越性.

关键词 生物信息学;类别不平衡;非编码 RNA 识别;SNP 位点鉴别;分类

中图法分类号 TP18

收稿日期:2009-05-11;**修回日期**:2009-10-12

基金项目:国家自然科学基金项目(60741001,60871092,60932008);黑龙江省杰出青年科学基金项目(JC200611);黑龙江省自然科学基金重点项目(ZJG0705)

0 引 言

分类问题是机器学习领域研究的重要课题之一,目前主要的机器学习方法处理的数据大都是各类样本数目相差不悬殊.如果训练集中的正反例样本数目相差悬殊,则会使得分类器的性能下降,通常会出现把整个样本空间都划为大类的情况[1],因此类别不平衡学习正受到越来越广泛的关注.

类别不平衡问题存在于许多领域,如金融欺诈检测[2]、石油勘探[3]、反垃圾邮件[4]等.普通机器学习的分类方法无法直接应用于这些领域.在生物信息学研究中,存在大量类别不平衡的分类问题.由于大多数问题中的正例来自于实验验证,而反例通常不需要实验验证,因此获取反例的成本低、正例的成本高,从而在训练集中通常出现反例远远多于正例的情况,比如:非编码 RNA 基因挖掘[5],尤其是 microRNA 挖掘[6].此外,在 SNP 位点判别[7]、microArray 数据分析[8]中也经常遇到这类问题.

在处理类别不平衡学习问题上,最早是使用随机采样的办法来更改训练集的样本,从而使训练集达到平衡.最简单的两种方法是随机过采样(over-sampling)和随机降采样(under-sampling).前者通过复制一些小类别的样本使数据集达到平衡,后者则随机选取大类中的一个子集以达到同样的目的.研究表明随机过采样的方法通常会带来时间开销大、过拟合等问题,因此目前主要采取的都是降采样的办法.但是降采样方法只使用了大类的一个子集,并没有充分利用已有的信息.在经历随机采样方法之后,出现了许多人工采样的方法.SMOTE[9]发展了过采样的思想.虽然也是增加小样本的数量,但是其增加的手段是通过人工生成,而不是直接随机选择复制,从而避免了过拟合问题,但有可能会引入噪声.类似地,也有通过人工手段进行降采样,从而有选择性地去除大类样本,达到类别平衡[10].

除了采样策略外,还有一些其他的策略也被应用于处理类别不平衡数据,如集成学习中的 Boosting 方法[11]、代价敏感学习算法[12]、单类学习方法(one class learning)[13]、级联的神经网络[14]、聚类方法和支持向量机[15]等.目前的 Boosting 方法由于反复训练而加大时间开销,无法应用于大规模数据处理.而代价敏感学习方法和单类学习方法都被证明等价于采样方法.与其类似的还有在自然语言理解领域获得了良好效果的基于聚类和支持向量机的方法.级联的神经网络虽然在 UCI 的部分数据集上取得了成功,但神经网络固有的随机性决定了其无法应用于更广泛的领域.

与其他应用领域不同,生物信息分类问题的属性通常全部是连续的,而且大部分是只有 2 类的判别问题.当分类属性连续的样本时,支持向量机(support vector machine, SVM)效果通常好于决策树等分类方法.另外,通用软件 LibSVM 具有使用简单、执行效率高等特点,因此被广泛地应用于生物信息学的分类问题中[16-17].然而 LibSVM 的参数调整主要依据准确率是否提高,因此 LibSVM 在类别不平衡样本的分类中,通常会把所有的测试样本全部划分为反例.目前的相关研究一般是在测试样本的反例中随机降采样,以达到正反例平衡.这种做法丢失了大部分反例样本信息,从而降低了分类器的性能.针对以上问题,提出了一种新方法处理类别不平衡的分类问题.

1 处理类别不平衡的分类方法

为了有效利用反例数据,同时避免训练时数据不平衡,本节首先提出了基于投票机制的集成学习方法.为了避免集成学习中弱分类器对最后投票结果的影响,本节又提出了基于重复训练错分样本的优化策略.最后讨论了类别不平衡对度量方法的影响.

1.1 基于投票机制的集成学习方法

为了弥补降采样丢失反类信息的特点,同时避免反例远大于正例的不平衡现象,本方法将反例随机等分若干份,每一份的大小与正例近似相等.然后用正例分别同每一组反例结合构成一组训练集.之后用不同的分类器对不同的训练集训练,从而得到一组分类器,这组中每一个分类器称为基分类器.当预测新的样本时,由这一组分类器投票产生预测结果.如算法 1 所示.

算法 1. 基于投票机制的类别不平衡分类训练算法.

输入:正例集 P,反例集 N ($|P| \ll |N|$);

输出:分类器 $F(x)$, x 是待测样本.

① $num \leftarrow \left\lfloor \frac{|N|}{|P|} \right\rfloor$;

② 随机重排 N,并将 N 等分 num 份,每一份记为 N_i;

③ for $i\in\{1,2,\cdots,num\}$

④　新建训练集 $T_i\leftarrow P+N_i$;

⑤　用 T_i 训练出分类器 C_i, $C_i(x)$ 表示对样本 x 的预测结果(1 为正例,−1 为反例);

⑥ end for

⑦ $F(x)=\operatorname{sgn}\sum_{i=1}^{num}C_i(x)$.

在训练每一个训练集时,本文采用原理不同的分类方法.因为 Krogh 等人研究表明:基分类器的差别越大,集成后的效果越好[18].本文调用了怀卡托智能分析环境 3.5(Waikato environment for knowledge analysis, WEKA)[19] 中可以处理实值属性的 38 种不同的分类器,包括决策树、随机森林、支持向量机、朴素贝叶斯、贝叶斯信念网、k 近邻等.当 $num\leqslant 38$ 时,我们随机取样生成一个测试集来评估这 38 种分类器,选取其中最好的 num 个分类器来训练;当 $num>38$ 时,38 种分类器全部被应用于训练集并且被循环利用,即在处理第 i 个训练集时,使用第 $i\%38$ 种分类器.

由于各个分类器没有进行参数性能优化,在处理分布较为复杂的样本时,大多数分类器都是弱分类器.在投票时,如果弱分类器过多,通常会影响集成后的结果.因此机器学习研究者提出了 AdaBoost 的思想,使得多个弱分类器集成后成为强分类器.然而在本文的学习方法中,对于每个分类器都采用 AdaBoost 策略,显然时间开销太大;如果对整个的训练集使用 AdaBoost 策略,集成的分类器会出现收敛过慢、甚至震荡的现象.因此,本文受 AdaBoost 思想启发,结合多训练集的特点,提出一种新的优化策略,以提高分类器的性能.

1.2　基于重复训练错分样本的优化策略

AdaBoost 通过增加错分类的样本数量、提高错分类样本的权重来使得多次训练后投票结果利于错分类的样本.然而迭代训练不适合用在大样本的数据集上,而在类别不平衡的分类问题中,反例集往往数据量很大.因此,本文受 AdaBoost 启发,利用投票过程中多数战胜少数的机制,多次训练错分类的样本,以期在集成后通过投票产生利于这些样本的预测结果.

在算法 1 中,由于是依次训练每一个分类器,因此在用 T_i 训练之后,可以用得到的基分类器 C_i 测试 T_i,从而得到被错分的样本集合 M_i. AdaBoost 的思想是将 M_i 再次加入到 T_i 中训练,得到 C_i' 和 M_i',循环执行,因而时间花销大.在本文中最后的结果已经是集成的结果,如果基分类器再是集成的结果,则是"集成的集成",显然浪费了资源.针对这一特点,本文提出:将 M_i 加入到 T_{i+1} 和 T_{i+2} 中.如果 M_i 中的某个样本 M_{ij} 已经存在于待加入的训练集中,则将其向后添加入下一个不包含它的训练集中.若已达到最后一个训练集时($i=num$),则将 M_{num} 加入到 T_1 和 T_2 中,重新训练直到某两个连续的分类器完全分类正确(当训练样本较小时)或准确率均超过一定的阈值(当训练样本较大时).训练过程如算法 2 所示.该策略可以替换算法 1 的步骤③～⑥.

算法 2. 基于重复训练错分样本的优化策略.

① $i\leftarrow 0$;

② $time\leftarrow 0$;

③ repeat

④ $i\leftarrow(i+1)\%num$;

⑤ 新建训练集 $T_i\leftarrow P+N_i$;

⑥ 用 T_i 训练出分类器 C_i, $C_i(x)$ 表示对样本 x 的预测结果(1 为正例,−1 为反例);

⑦ 用 T_i 测试 C_i, M_i 表示错误样本集合;

⑧ for $j\in\{1,2,\cdots,|M_i|\}$

⑨ $first\leftarrow i$;

⑩ repeat

⑪ $first\leftarrow(first+1)\%num$;

⑫ until T_{first} 中没有 M_{ij};

⑬ $T_{first}\leftarrow M_{ij}$;

⑭ $T_{(first+1)\%num}\leftarrow M_{ij}$;

⑮ end for

⑯ if $i:=0$

⑰ $time\leftarrow time+1$;

⑱ until $time=max\_repeat\_times$ 或 M_i, M_{i-1} 均为空.

本策略之所以要把错分的样本加入到后续两个训练集中,是因为如果只加入到下一个训练集中,无法保证最终投票时对于该样本正确分类的分类器能够胜出错分类的分类器.如果加入到多于两个的训练集中,又增添了计算开销.当把错分的样本加入到下两个训练集时,即使是又出现错分情况,其又被加入到接下来的两个不包含它的训练集,直到它被连续的两个分类器正确分类.这种做法能保证对其正确分类的分类器的个数比错分的分类器的个数多 1.因为该策略如同一个满二叉树的生成过程,错分

的分类器永远都是内部节点,它的下两个分类器就是它的两个儿子,而正确分类的分类器是叶子,由图论的基本知识,满二叉树的叶子节点比内部节点多1.因此可以保证在最后投票时,该样本可以被正确的分类.图1的例子是一个样本添加过程和满二叉树的对应关系.

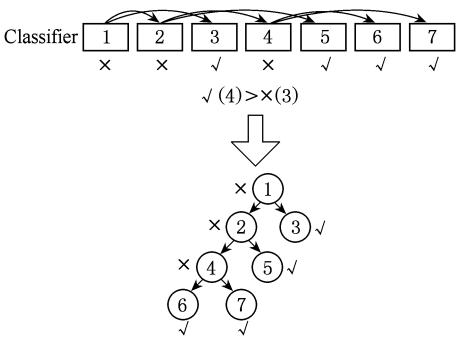

Fig. 1 An example for algorithm 2.

图1 算法2一例示意图

本文的做法避免了基分类器中的集成,从而节省了时间开销.尽管也循环训练,但实验证明基于该策略的训练时间仅仅略高于算法1,远远小于对基分类器或集成后的分类器使用AdaBoost.

1.3 类别不平衡的评价方法研究

在处理类别不平衡数据时,衡量分类器的性能指标也与平时有所差异.

生物信息学中的预测(分类)问题一般用敏感性 sn(sensitivity)、特异性 sp(specificity)、准确率 ACC(overall accuracy)和马修兹系数 MCC(Matthew's correlation coefficient)综合来衡量分类器的效果.如果用 TP 表示正确预测到的正例的个数,TN 表示正确预测到的反例的个数,FP 表示把反例预测成正例的个数,FN 表示把正例预测成反例的个数,那么则有:

$$sn = \frac{TP}{TP + FN}; \tag{1}$$

$$sp = \frac{TN}{FP + TN}; \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}; \tag{3}$$

$$MCC = \left\{ TP \times TN - FP \times FN \Big/ \sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)} \right\}. \tag{4}$$

敏感性 sn 即相当于模式识别中的"查全率" $recall$,而模式识别中的"查准率" $precision$ 与准确率 ACC 略有差异,定义为

$$precision = \frac{TP}{FP + TP}. \tag{5}$$

当样本类别平衡时,一般情况下 MCC 是介于 sn 和 sp 中间的某个值,它可以综合衡量软件的效果,因为它既考虑了敏感性也考虑了特异性.然而当样本类别不平衡时,按照 MCC 的定义,MCC 的值通常远小于 sn 和 sp.因为此时,TN 和 FP 是一个数量级,远远大于 TP 和 FN.而

$$MCC \approx \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP} \times \frac{TN+FP}{FP} \times \frac{TP}{TP+FP}} \approx \sqrt{sn \times sp} \times \sqrt{\frac{TP}{TN}} \times \sqrt{\frac{TN+FP}{FP}} = \sqrt{sn \times sp} \times \sqrt{\frac{TP}{TN}} \times \frac{1}{\sqrt{1-sp}}. \tag{6}$$

在类别不平衡问题中,$\sqrt{\frac{TP}{TN}}$ 受类别样本比例的影响,通常会非常小,远远小于 $\sqrt{sn \times sp}$ 和 $\frac{1}{\sqrt{1-sp}}$,此时 MCC 主要受样本比例影响,而无法真实地反映分类的效果.

另外,在类别不平衡分类问题中,由于 TN 和 FP 远大于 TP 和 FN,那么

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \approx \frac{TN}{TN + FP} = sp. \tag{7}$$

因此 ACC 通常近似等于 sp,所以在生物信息学中衡量类别不平衡的分类效果时,一般用 sn 和 sp 即可,不用进一步考虑 MCC 和 ACC.

2 实验结果与分析

为了验证本文提出的方法和策略的有效性,本节分别利用UCI测试数据和生物信息学中的真实数据对本文方法进行验证.实验设计思路如下:首先选择UCI中的类别不平衡数据,同目前处理类别不平衡的主要方法进行对比,来验证本算法的有效性.同时对比不使用和使用重复训练错分样本优化策略的差异,来证明该策略的优越性.然后在两个不同的生物信息学问题(snoRNA和microRNA前体的判别)中测试,并和已发表的成果进行对比,以展示我

们的方法对实际研究的提高.最后将其应用于一个最新的生物信息学研究(在 EST 序列中挖掘 SNP 位点)中,取得了良好的效果并解决了实际问题.

2.1　UCI 数据

本文选取 cmc, haberman, ionosphere, letter 和 pima 5 组 UCI 测试数据,这 5 组数据的特点是属性均为实数,样本类别不平衡(本文将多类问题中最小的类别视为正例,其余的视为反例).与本文对比的分类方法分别有 AdaBoost(弱分类器为决策树方法)、随机降采样(UnderSampl)、混合采样(HSampl)、AsymBoost 和 BalanceCascade.这 5 种方法在这 5 组 UCI 测试数据上的实验结果来自文献[14].

此外,为了验证重复训练错分样本优化策略的效果,我们对比了 1 次训练和循环训练 2 种状态本文方法的分类效果.循环训练指算法 2 中最后 1 个分类器的错分数据再添加入第 1,2 个分类器的训练集中,重新训练分类器直到有分类器不再错分训练数据为止.而 1 次训练是指最后 1 个分类器的错分数据不再向其他分类器的训练集添加,当所有的分类器训练过 1 遍之后就开始对测试数据进行投票表决.注意这里的 1 次训练,同样把 1 个分类器的错分数据添加入后续 2 个分类器的训练集,只是没有循环训练.

如表 1 所示,在 5 组 UCI 测试集中,只有在 letter 数据集上我们的软件性能略差,其他 4 组本文方法的性能均优于其他的分类方法.我们的方法是以集成学习理论为基础,因此更适合处理弱分类的数据集(如 cmc, haberman),而 letter 是一个强可分的数据集,特别是基于决策树的 AdaBoost 已经达到几乎完全准确的程度,而我们的方法受到其他分类器的影响,分类效果稍弱.但总的来说,尤其是处理类别不平衡的弱分类问题,我们的方法显现了较大的优势.尽管我们多次使用正例集,实验表明通过集成的机制,并没有损害到查准率,却有效地提高了查全率.

另外,在 5 组实验数据上,有 3 组数据(haberman, ionosphere, pima)的循环训练的效果好于一次训练,有 1 组略逊(letter),另 1 组(cmc)在查准率和查全率上互有胜负.letter 略逊是由于该组数据是强分类数据,每个分类器的错分数据都非常少,循环训练意义不大.因此我们可以看出,循环训练错分类的训练样本,可以提高弱分类器的集成效果.但循环训练同时也增大了时间开销,在数据规模较小时,这种开销可以忽略,但当样本个数较多、特征空间复杂的时候,循环训练的时间可能是 1 次训练 2 倍,甚至更多.

Table 1　Performance of 7 Different Classifiers on 5 UCI Data Sets

表 1　7 种分类器在 5 个 UCI 数据集上的表现

Data(|*P*|/|*N*|)	Classifier	*precision*	*recall*
cmc(333/1140)	AdaBoost	0.40	0.39
	UnderSampl	0.33	0.63
	HSampl	0.37	0.48
	AsymBoost	0.39	0.42
	BalanceCascade	0.35	0.59
	LibID(once)	0.48	0.74
	LibID(repeat)	0.50	0.67
haberman(81/225)	AdaBoost	0.35	0.36
	UnderSampl	0.36	0.60
	HSampl	0.36	0.47
	AsymBoost	0.34	0.39
	BalanceCascade	0.36	0.57
	LibID(once)	0.54	0.80
	LibID(repeat)	0.59	0.84
ionosphere(126/225)	AdaBoost	0.95	0.88
	UnderSampl	0.92	0.89
	HSampl	0.94	0.86
	AsymBoost	0.95	0.88
	BalanceCascade	0.93	0.89
	LibID(once)	0.94	0.89
	LibID(repeat)	0.94	0.91
pima(268/500)	AdaBoost	0.63	0.60
	UnderSampl	0.58	0.73
	HSampl	0.62	0.65
	AsymBoost	0.63	0.61
	BalanceCascade	0.60	0.71
	LibID(once)	0.78	0.76
	LibID(repeat)	0.77	0.81
letter(789/19211)	AdaBoost	0.99	0.98
	UnderSampl	0.83	0.99
	HSampl	0.92	0.99
	AsymBoost	0.99	0.98
	BalanceCascade	0.96	0.99
	LibID(once)	0.88	0.99
	LibID(repeat)	0.85	0.98

Note: Data in this table are average values of 10 times 5 cross-validation.

2.2 识别 snoRNA

核仁小分子 RNA(snoRNA)是一种重要的非编码 RNA,它可以指导核糖体 RNA(rRNA)的甲基化和假尿嘧啶化,进而影响其生物合成.另外它还可以指导小核 RNA(snRNA)、转运 RNA(tRNA)和信使 RNA(mRNA)的转录后修饰.根据结构特点,snoRNA 主要可以分为 C/D box snoRNA 和 H/ACA box snoRNA 两大类.

Jana 等人的研究表明:2 种不同的 snoRNA 在二级结构、自由能、GC 含量、配对碱基个数等特征上相对于随机的基因组序列均具有显著性,因此可以用分类的方法从众多的非编码 RNA 中找出 C/D box snoRNA 和 H/ACA box snoRNA[17].

对于 C/D box snoRNA,Jana 等选取了 306 个正例和 45 209 个反例作为训练集;对于 H/ACA box snoRNA,Jana 的训练集中有 65 个正例和 8 445 个反例.他们使用 LibSVM 作为分类器.在这 2 个训练集上,我们使用与文献[17]中同样的特征,表 2 是 LibSVM 和本文算法在 5 重交叉验证上的实验结果对比.

Table 2 Performance of LibSVM and Our Method on snoRNA

表 2 本文方法和 LibSVM 在 snoRNA 上的效果比较

RNA	Measurement	LibSVM	LibID
H/ACA box snoRNA	*sn*	0.78	**0.86**
	sp	0.89	**0.90**
C/D box snoRNA	*sn*	0.96	**0.90**
	sp	0.91	**0.94**

由表 2 可以看出,对于弱分类问题 H/ACA box snoRNA,我们的方法无论是敏感性 *sn* 还是特异性 *sp* 都有显著的提高.对于强分类问题 C/D box snoRNA,我们的方法在保证较高的敏感性的同时,提高了特异性.这对于分子生物学研究者是非常重要的,因为生物学实验验证的成本非常高,因此一般对生物信息预测软件的特异性要求高于敏感性.

2.3 判别 microRNA 前体真伪

microRNA 是生物体内另外一种重要的非编码 RNA 分子,在调解遗传基因表达、控制细胞生长等方面有着重要的作用.在各种生物基因组中寻找 microRNA 是诠释基因组工作的一个重要的部分,其思路是在基因组序列中找出可疑的片段然后鉴别.目前鉴别的方法主要是生物芯片(microArray)或北桥实验(Northern Blot),它们都具有花费高、操作困难和不完全准确的缺点.因此生物信息学研究者试图通过机器学习的方法来对其分类.

由于 microRNA 的成熟体较短,不容易判别,因此一般对其前体(precursor)提取二级结构特征,从而进行判别.然而目前用实验确定的 microRNA 只有几千个,在一个物种上的则更少,而类似于前体的发夹环则可以在基因组中找到很多,对于人的基因组至少可以找到上百万条.因此这是一个明显的类别不平衡的分类问题.Xue 等人[16]对人类的 microRNA 前体进行了研究,他们提供的数据集中存在 193 个正例、8 494 个反例,而在使用 LibSVM 时通过随机降采样,提取了 163 个正例和 168 个反例作为训练集,用 30 个正例和 1 000 个反例作为测试集.我们选用了和他们相同的测试集,而训练集则使用了除测试集以外的所有样本(163 个正例、7 494 个反例),表 3 是实验结果对比,其中 Triplet-SVM 是 Xue 等人[16]提供的软件.

Table 3 Performance of Our Method and Triplet-SVM on miRNA

表 3 与 Triplet-SVM 的效果比较

Measurement	Triplet-SVM	LibID
sn	0.93	**0.83**
sp	0.88	**0.91**

本文的方法更多地考虑了反例信息,因此 *sp* 要高于 Triplet-SVM.而 Triplet-SVM 中的 *sn* 高于本文算法的结果,是由于其训练集的正例远高于测试集,因此存在"过拟合"的现象.这一点在 Xue 等人[16]的论文中也被提及,当他们用同样的训练集去预测其他物种时,*sn* 有所下降.另外,同 C/D box snoRNA 的分类结果一样,本文提出的分类器在保证 *sn* 的情况下提高了 *sp*,这对于分子生物学研究人员是非常重要的.

Xue 等人的主要贡献在于特征提取,通过选择合适的特征使得其分类器成为强分类器.而本文的工作是基于集成学习和 AdaBoost 思想,因此更适合处理弱分类的问题,比如在 EST 序列中判别真实的 SNP 位点.

2.4 EST 序列中挖掘 SNP 位点

SNP 位点是重要分子标记手段,许多研究表明 SNP 同人群分类、遗传疾病都有着紧密的联系.在 EST 序列中挖掘 SNP 位点,进而进行分子标记,是一项可以节省大量实验成本却又富有挑战性的任务.

首先在人类的部分 EST 序列(22 994 条)中,利

用多序列比对的办法,找到了 3 074 个候选的 SNP 位点.通过与 NCBI dbSNP 数据库比较,确定了其中有 183 个真实的 SNP 位点.由于反例样本(2 891 个)远远大于正例样本(183 个),无法直接用 LibSVM 进行处理.第 1 次实验用降采样的方法结合 LibSVM,第 2 次实验用类似于本文的分割反例集然后投票的方法,基分类器使用 LibSVM.表 4 是 2 次实验与本文算法的效果对比.

由表 4 可以看出,投票机制优于降采样机制.在投票机制下,利用多种分类器且重复训练错分样本的本文方法的效果好于仅使用 LibSVM.因此本实验证明了本文使用的 3 个主要策略的优越性:1.分割投票策略;2.使用原理不同的基分类器策略;3.循环训练错分样本策略.

Table 4　Performance of LibSVM and Our Method on SNP Data
表 4　与 LibSVM 的效果比较

Measurement	LibSVM(Under-Sampling)	LibSVM(Voting)	LibID
sn	0.50	0.66	**0.81**
sp	0.69	0.70	**0.82**

Note: Data in this table are average value of 10 times 5 cross-validation.

3　结束语

为了处理生物信息学中的样本类别不平衡问题,本文提出了一种基于分割反例集并投票的决策方法.在处理强分类问题时,能够在保证敏感性的同时,提高特异性,这对于生物信息研究者十分重要.在生物信息学研究中,特异性往往比敏感性重要,因为高特异性可以降低实验验证成本.

不平衡数据的分类问题是一个很重要的课题.本文的方法仅应用于生物信息学中常见的几个挖掘问题,对基因芯片这种高维极度不平衡数据的分析尚需要进一步的研究.另外,分类器的效果主要受数据分布的影响,在考虑数据分布特点的同时,研究不平衡程度对该方法的影响将是未来的工作.本文数据、软件的下载地址为 http://nclab.hit.edu.cn/zouquan/libid/.

参　考　文　献

[1] Xu Yan, Li Jintao, Wang Bin, et al. A study of feature selection for text categorization on imbalanced data [J]. Journal of Computer Research and Development, 2006, 43 (Suppl): 58-62 (in Chinese)
(徐燕,李锦涛,王斌,等. 不均衡数据集上文本分类的特征选择研究[J]. 计算机研究与发展, 2006, 43(增刊): 58-62)

[2] Stolfo S, Fan W, Lee W, et al. Cost-based modeling for fraud and intrusion detection: Results from the jam project [C] //Proc of the 5th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1999

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

[4] Fawcett T. "In vivo" spam filtering: A challenge problem for data mining [J]. ACM SIGKDD Explorations, 2003, 5(2): 140-148

[5] Wang Chunlin, Ding Chris, Meraz R F, et al. PSol: A positive sample only learning algorithm for finding non-coding RNA genes [J]. Bioinformatics, 2006, 22(21): 2590-2596

[6] Jiang P, Wu H, Wang W, et al. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features [J]. Nucleic Acids Research, 2007, 35: W339-W344

[7] Marth G T, et al. A general approach to single-nucleotide polymorphism discovery [J]. Nature Genetics, 1999, 23(4): 452-456

[8] Li Jianzhong, Yang Kun, Gao Hong, et al. Model-free gene selection method by considering unbalanced samples [J]. Journal of Software, 2006,17(7): 1485-1493 (in Chinese)
(李建中,杨昆,高宏,等. 考虑样本不平衡的模型无关的基因选择方法[J]. 软件学报, 2006, 17(7): 1485-1493)

[9] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(6): 321-357

[10] Batista G E, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations, 2004, 6(1): 20-29

[11] Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach [J]. ACM SIGKDD Explorations, 2004, 6(1): 30-39

[12] Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting [C] //Proc of the 3rd Int Conf on Data Mining. Piscataway, NJ: IEEE, 2003: 435-442

[13] Manevitz L M, Yousef M. One-class SVMS for document classification [J]. Journal of Machine Learning Research, 2001, 2(2): 139-154

[14] Liu Xuying, Wu Jianxin, Zhou Zhihua. A cascade-based classification method for class-imbalanced data [J]. Journal of Nanjing University: Natural Sciences, 2006, 42(2): 148-155 (in Chinese)
(刘胥影,吴建鑫,周志华. 一种基于级联模型的类别不平衡数据分类方法[J]. 南京大学学报:自然科学, 2006, 42(2): 148-155)

[15] Li Peng, Wang Xiaolong, Liu Yuanchao, et al. A classification method for imbalance data set based on hybrid strategy [J]. Acta Electronica Sinica, 2007, 35(11): 2161-2165 (in Chinese)
(李鹏, 王晓龙, 刘远超, 等. 一种基于混合策略的失衡数据集分类方法[J]. 电子学报, 2007, 35(11): 2161-2165)
[16] Xue C, Li F, He T, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine [J]. BMC Bioinformatics, 2005, 6: 310
[17] Hertel J, Hofacker I L, Stadler P F. snoReport: Computational identification of snoRNAs with unknown targets [J]. Bioinformatics, 2008, 24(2): 158-164
[18] Krogh A, et al. Neural network ensembles, cross validation, and active learning [G] //Advances in Neural Information Processing Systems 7. Cambridge: MIT Press, 1995: 231-238
[19] Frank E, et al. Data mining in bioinformatics using Weka [J]. Bioinformatics, 2004, 20(15): 2479-2481

Zou Quan, born in 1982. PhD. His main research interests include the prediction of ncRNA structure and mining ncRNA.
邹 权, 1982 年生, 博士, 主要研究方向为非编码 RNA 的结构预测与挖掘算法.

Guo Maozu, born in 1966. PhD. Professor since 2002. PhD supervisor. His main research interests include bioinformatics and machine learning.
郭茂祖, 1966 年生, 博士, 教授, 博士生导师, 主要研究方向为生物信息学与机器学习(maozuguo@hit.edu.cn).

Liu Yang, born in 1976. PhD and lecturer since 2006. His main research interests include machine learning and computer vision
刘 扬, 1976 年生, 博士, 讲师, 主要研究方向为机器学习和计算机视觉.

Wang Jun, born in 1983. PhD candidate. Her main research interests include algorithms and application on SNP, analysis of disease associations.
王 峻, 1983 年生, 博士研究生, 主要研究方向为 SNP 分析算法与应用、疾病关联性分析.

Research Background

Two-class classification is usually adopted for mining or predicting the bioinformatics data. However, as the distribution of the native data or the experiment cost, the training data of one class (negative set) often outnumbers the other class(positive set). In this paper, we propose a novel ensemble learning method and develop the software program LibID for the imbalance data. It works well on some bioinformatics problems, including microRNA precursors and snoRNA identification, mining SNP sites from aligned ESTs.

This work is supported by the National Natural Science Foundation of China under grant No. 60671011, No. 60741001, No. 60871092 and No. 60932008, the Science Fund for Distinguished Young Scholars of Heilongjiang Province in China under grant No. JC200611, and the Natural Science Foundation of Heilongjiang Province in China under grant No. ZJG0705.