一种基于聚类提升的不平衡数据分类算法

胡小生[1] 张润晶[2] 钟 勇[1]

[1](佛山科学技术学院电子与信息工程学院 佛山 528000)
[2](佛山科学技术学院信息与教育技术中心 佛山 528000)

摘 要 不平衡数据分类是机器学习研究领域中的一个热点问题。针对传统分类算法处理不平衡数据的少数类识别率过低问题，文章提出了一种基于聚类的改进 AdaBoost 分类算法。算法首先进行基于聚类的欠采样，在多数类样本上进行 K 均值聚类，之后提取聚类质心，与少数类样本数目一致的聚类质心和所有少数类样本组成新的平衡训练集。为了避免少数类样本数量过少而使训练集过小导致分类精度下降，采用少数过采样技术过采样结合聚类欠采样。然后，借鉴代价敏感学习思想，对 AdaBoost 算法的基分类器分类误差函数进行改进，赋予不同类别样本非对称错分损失。实验结果表明，算法使模型训练样本具有较高的代表性，在保证总体分类性能的同时提高了少数类的分类精度。

关键词 不平衡数据分类；K 均值聚类；AdaBoost；集成学习
中图分类号 TP 18 **文献标志码** A

A Clustering-Based Enhanced Classification Algorithm for Imbalanced Data

HU Xiaosheng[1] ZHANG Runjing[2] ZHONG Yong[1]

[1](*College of Electronic and Information Engineering*, *Foshan University*, *Foshan* 528000, *China*)
[2](*Information and Education Technology Center*, *Foshan University*, *Foshan* 528000, *China*)

Abstract Imbalanced data exist widely in the real world and their classification is a hot topic in the field of machine learning. A clustering-based enhanced AdaBoost algorithm was proposed to improve the poor classification performance produced by the traditional algorithm in classifying the minority class of imbalanced datasets. The algorithm firstly constructs balanced training sets by the clustering-based undersampling, using K-means clustering to cluster the majority class and extract cluster centroids and then merge with all minority class instances to generate a new balanced training set. To avoid the declining of the classification accuracy caused by the shortage of training sets owing to too few minority class samples, SMOTE (Synthetic Minority Oversampling Technique) combining the clustering-based undersampling was used. Next, the misclassification loss function in the basic classifier of the AdaBoost algorithm was modified based on the cost-sensitive learning theory to assign asymmetric misclassification losses to samples of different classes. The experimental results show that, the proposed algorithm makes the model training samples more representative and greatly increases the classification accuracy of the minority class, keeping the overall classification performance.

Keywords imbalanced data classification; K-mean clustering; AdaBoost; ensemble learning

收稿日期：2013-08-22
基金项目：广东高校优秀青年创新人才培养项目(2013LYM_0097)；佛山市智能教育评价指标体系研究(DX20120220)；佛山科学技术学院校级科研项目。
作者简介：胡小生(通讯作者)，硕士，讲师，高级工程师，研究方向为机器学习和数据挖掘，E-mail：feihu@fosu.edu.cn ；张润晶，高级工程师，研究方向为信息检索和信息安全；钟勇，博士，教授，研究方向为信息安全、信息检索和云计算。

1 引 言

不平衡数据集是指在一个数据集中，某些类的数量远远大于其他类别的数量，其中类别数量多的为多数类，类别数量少的为少数类。在现实应用领域中，广泛存在着不平衡数据集：文本分类、医疗诊断、信用卡诈骗检测和网络入侵检测等，在处理这些情况的过程中，少数类的识别准确率更为重要，其错分代价更大。传统分类方法为保证总体分类精度，通常将少数类误分到多数类来保证整体分类精度，实际分类效果并不理想。因此，如何有效地对不平衡数据进行分类是当今机器学习和数据挖掘领域研究的一个热点问题。

鉴于不平衡数据分类的重要性，国内外学者进行了大量研究，现有的不平衡数据处理方法主要有两个方面：

(1)数据层面，改变数据的分布。最简单的两种方法是随机过采样(Oversampling)和随机欠采样(Undersampling)，前者对少数类样本复制使数据分布相对平衡，后者通过抽取一部分多数类样本达到数据平衡目的。两者各有缺点：过采样通过不断复制少数类而使数据规模变大，使分类器学习到的决策域变小，从而容易导致过拟合的问题；欠采样由于抽取部分多数类样本使信息丢失严重。目前，很多学者提出改进的数据采样方法[1-5]。为了避免随机过采样的不足，Chawla等[1]提出一种少数过采样技术(Synthetic Minority Oversampling Technique，SMOTE)算法，通过采用少数类样本合成技术产生新的样本，该算法可使少数类具有更大的泛化空间，但也可能导致分类器过于拟合，同时不可避免地会产生噪音样例或者边际样例。基于此，Batista 等提出了 SMOTE 与 Tomek links[2]相结合的数据平衡方法[3]。Yen 等[4]提出了一种基于聚类的抽样方法 SBC(Undersampling Based on Clustering)。SBC 通过聚类后簇内的多数类与少数类的比例确定抽样参数，但该算法忽略了数据分布的特征，导致样本代表性差，不能反映原始数据的分布。蒋盛益等[5]提出了一趟聚类的数据下抽样算法，根据训练样本聚类后簇的特征与数据倾斜程度确定抽样比例，该方法较好地保持了少数类信息，缩小了数据分布的差异，提高了分类的性能，但该方法面临着如何自适应确定抽样比例参数以对不同密度的簇分离的问题。

(2)算法层面，修改已有的分类算法或者提出新的算法。代价敏感学习、主动学习、集成学习以及单类别学习等，是处理不平衡数据集的常见算法[6-8]。其中，代价敏感学习赋予各个类别不同的错分代价，研究表明代价敏感学习和不平衡数据学习之间存在很强的联系，代价敏感学习的相关理论和算法可以用来解决不平衡数据的学习问题[9]。集成学习通过对多个分类器的分类识别结果进行融合能很好地提高单一目标的分类识别效果，作为集成学习方法的 boosting 提升技术用于提高分类性能，无论数据集是否平衡，都可以通过 boosting 迭代创建集成模型，提升弱分类器的性能。当前，将 boosting 技术应用于不平衡数据分类主要有两类：一种将代价敏感学习和 boosting 技术相结合，例如 AdaCost[10]和 RareBoost[11]；另一类是将数据采样处理方法和 boosting 技术相结合，例如 SMOTEBoost[12]。

本文提出一种融合无监督聚类和 boosting 提升技术的不平衡数据分类算法——基于聚类改进 AdaBoost 分类算法。为了研究方便，本文主要关注二分类情况，少数类也称为正类，多数类称为负类，而正类、负类的类别标签取值分别为{＋1，－1}。算法首先进行基于聚类的欠采样，在负类样本上进行 K 均值聚类，使聚类数量与正类样本数量相同。之后每个聚类得到聚类质心，将所有的聚类质心与正类样本组成平衡的训练集，参与后续改进 AdaBoost 算法的训练。最后借鉴代价

敏感学习思想，对 AdaBoost 错分的正类样本赋予更大的错分代价，进而修改了各个基分类器的输出决策权重，最终得到分类集成学习模型。

2　基于聚类的数据欠采样

对于不平衡数据集，对负类样本进行欠采样和在正类进行过采样均能改变数据分布，使数据达到平衡。但仍存在缺点：过采样容易导致过度拟合问题，欠采样则会引起信息丢失。为了抽取最具代表的训练样本，需对样本进行划分。本文选取 K 均值聚类方法进行欠采样，将训练集中的负类样本聚类为 k 个不相交的子集，然后，在各子集上提取最富有代表性的样本信息，与正类样本组成新的平衡训练集。

本方法首先提取训练集中的正类样本个数 k，并以 k 为聚类中心数目对训练集中所有的负类样本进行 K 均值聚类，提取 k 个聚类质心。“负”类样本的 k 个聚类质心加上所有的正类样本组成一个新的平衡训练集。

获取聚类质心的具体流程如下：

算法 1　聚类欠采样算法：

输入：数据集 D

输出：平衡的训练集 D_{train}

(1) S=pre_process[D]；// 数据预处理

(2) train[S]＝随机取数据集 S 中的 80% 样本，test[S]＝S 中剩余的 20% 样本；

(3) 提取 train[S] 中正、负类样本集合 S^+、S^-，计算 S^+中正类样本数量 k，令 K=k；

(4) 对集合 S^-的样本进行 K 均值聚类，得到 k 个不相交子集及其聚类质心；

(5) 提取 k 个聚类质心记为集合 S'；

(6) S'_{train}=$S^+ \cup S'$

由于数据集中的属性一般有连续型和分类型两种，因此，在算法 1 的第一步须进行数据预处理 pre_process() 过程。具体方法为：对于数据集中的分类型属性，采用二进制编码方式进行转换，将分类型的属性转换为若干个取值 0 和 1 的属性；而对于连续型属性，为了消除不同量纲所造成的影响，对输入数据进行最大最小归一化处理，进而将所有数据归一化到[0,1]之间，其公式如下所示：

$$a_j=\frac{a_j-Min_A}{Max_A-Min_A} \tag{1}$$

其中，Max_A、Min_A 分别表示特征 A 上数据的最大值和最小值。

算法 1 的实质是为了组成一个新的平衡数据集，对负类样本进行聚类压缩，由聚类的质心代表聚类的所有样本。此方法对大规模数据（正类样本数量比较大）能取得较好的效果，但若在正类样本数量很小时，单纯使用此方法将会导致输出的平衡训练集过小而难以到达理想的分类精度。针对此问题，提出一个将 SMOTE 与聚类欠采样相结合方法，在过采样与欠采样之间寻找一个平衡点。其中，SMOTE 通过生成合成样本对正类样本进行过采样。具体为：对于每个正类样本 x，在其同类中查找 n 个近邻，根据上采样的倍率 N，在样本 x 和被选中的近邻样本之间进行随机插值，生成新的样本。

3　AdaBoost 算法及其改进

3.1　AdaBoost 算法简介

AdaBoost 算法是一种典型的集成学习算法，可以有效提高单一学习器的泛化能力。它首先赋予训练集中每个训练样例相同的初始权重，然后通过若干轮训练得到若干弱分类器，在每一轮训练结束后，增加没有正确分类的样本的权值，减少正确分类的样本的权重，使系统在下一轮训练中更加关注那些分类错误的样本，最后这些弱分类器通过加权集成为一个强分类器完成分类任务。算法中每个基分类器的投票权值计算如下：

$$a_t=\log\frac{1-\varepsilon_t}{\varepsilon_t} \tag{2}$$

其中 ε_t 表示 t 轮分类器的分类误差，其定义为：

$$\varepsilon_t=\sum_{i=1}^{N}\omega_t(i)\cdot[\![h_t(x)\neq y_i]\!] \tag{3}$$

$\omega_t(i)$表示样本 i 在第 t 轮迭代时的权重，$\omega_1(i)=1/N$；$[\![\pi]\!]$表示条件 π 满足时输出 1，否则输出 0，如果 $\varepsilon_t=0$ 或 $\varepsilon_t\geqslant 0.5$ 则终止。

子分类器 h_t 形成后，实例样本的权重更新公式如下：

$$\omega_{t+1}(i)=\frac{\omega_t(i)\exp(-a_t y_i h_t(x))}{Zt} \tag{4}$$

Z_t 为归一化常数。最终输出：

$$H(x)=sign(\sum_{i=1}^{T}a_t h_t(x)) \tag{5}$$

3.2 AdaBoost 算法改进

在 AdaBoost 算法中，每个基分类器的投票权重是基于总体的误分情况，目的是减少平均误分率，也就是提高总体分类正确率。对于类别平衡的数据集来说，这种学习方法是可靠的。然而对于类别不平衡数据集，单纯地追求基分类器的分类精度，对合成分类器的分类效果影响并不直接。因此为了在总体上获得较高的精度，分离器通常倾向于忽视数量较少的正类样本，结果使得到的分类器在正类上识别效果差。而在实际中，正类样本的识别往往是最需要关注的。因此，本文考虑对 AdaBoost 算法中基分类器的投票权重进行改进，使其充分考虑到正类的样本数据。

改进 AdaBoost 算法基分类器的投票权重具体做法是对公式(3)中的 ε_t 的定义计算方式进行修改，进而改变了基分类器的输出投票权重 a_t。改进方法实质是借鉴代价敏感学习思想，对基分类器的误分代价在各个类别上不再一视同仁，对正类样本的错分，赋予更大的误分代价。假设原始训练集的不平衡度为 r，r＝负类样本数量/正类样本数量，则修改公式(3)：

$$\varepsilon_t=\sum_{i=1}^{N}r^{\frac{1+y_i}{2}}\cdot\omega_t(i)\cdot[\![h_t(x)\neq y_i]\!] \tag{6}$$

公式(6)与公式(3)相比，在计算分类器的误分代价时，对每个样本实例的误分代价乘上一个系数 $r^{\frac{1+y_i}{2}}$，当样本为负类时，$y_i=-1$，$r^{\frac{1+y_i}{2}}=1$ 样本误分代价与初始算法相比不变；当样本为正类时，$y_i=+1$，$r^{\frac{1+y_i}{2}}=r$ 样本误分代价需乘上系数 r。公式(6)表明，当某个基分类器误分较多正类样本时，其误分总代价 ε_t 增大，相应在最终决策输出时其投票权重 a_t 值变小。

4 实验结果及分析

4.1 数据集

通常情况下，将不平衡度在[1.5, 3.5)、[3.5, 9.5)、[9.5, ＋∞)分别称为低度不平衡范围、中度不平衡范围和高度不平衡范围。为了评估算法的性能，选择 8 组具有不同实际应用背景的不同平衡度的 UCI 数据，如表 1 所示。对于含有多个类别的数据，采用与其他文献相似的方法：将其中的一类作为少数类，合并剩下的各个类别成为一个整体为多数类。例如，将 page-blocks 的类别 5 作为少数类，合并其他的类作为多数类。

4.2 评价标准

在传统的分类学习中，一般采用分类精度(分类正确的样本个数占总样本个数的百分比)作为评价指标，然而对于不平衡数据集，这一指标实际意义不大，因为它反映的是多数类样本的分类测试结果。针对不平衡数据，很多学者提出建

表 1　UCI 数据集

Table 1.　UCI datasets

数据集	样例数目	少数类	多数类	不平衡度
pima_diabetes	768	268	500	1.87
waveform	5000	1655	3345	2.02
splice	3190	767	2423	3.16
artificial	5109	708	4401	6.21
optdigits	5620	562	5058	9.41
letter	20000	734	19266	26.25
nursery	12960	328	12632	38.51
page-blocks	5473	115	5358	46.59

立在混淆矩阵基础上的 *F-measure*、*G-mean* 等评价指标[13]，混淆矩阵如下表 2 所示。

表 2　混淆矩阵

Table 2.　Confusion matrix

样本类型	预测正类	预测反类
实际正类	TP	FN
实际反类	FP	TN

在某些应用中，人们更加关注少数类样本的分类性能，*F-measure* 就是用于衡量少数类分类性能的指标。*F-measure* 是查全率(*recall*)和查准率(*precision*)的调和均值，其取值接近两者的较小者，因此，较大 *F-measure* 值表示 *recall* 和 *precision* 都较大：

$$F-measure=\frac{2\times recall\times precision}{recall+precision} \tag{7}$$

其中：

$$recall=\frac{TP}{TP+FN},precision=\frac{TP}{TP+FP} \tag{8}$$

G-mean 是一种衡量分类器整体分类性能的评价指标，其定义如下：

$$G-mean=\sqrt{\mathrm{acc}^{+}\times acc^{-}} \tag{9}$$

其中：acc^{+}表示正类的分类精度，acc^{-}表示负类的分类精度。

从定义中可以看出，*G-mean* 兼顾了少数类和多数类精度的平均，在保持正、负类分类精度平衡的情况下最大化两类的精度，能够反映出分类器的整体性能。

本文采用 *F-measure* 和 *G-mean* 作为评价标准。其中，使用 *F-measure* 来衡量正类的分类性能，而使用 *G-mean* 来衡量整体分类性能。

4.3　实验结果

在 weka3.6.3 环境下对本文算法进行了验证，并且与传统的分类算法 AdaBoost、SMOTEBoost 和 RUSBoost 进行了比较，相关结果如表 3 和表 4 所示。实验中，AdaBoost、SMOTEBoost 和 RUSBoost 算法的基分类器均采用 J48 算法，本文所提算法的基分类器则采用在小样本平衡集上分类性能表现优异的支持向量机(Support Vector Machine)算法。为比较方便

起见，实验中对数据采用五折交叉验证(5-fold cross-validation)方式。为了保证数据在进行分组过程中不平衡度保持一致，采用分层采样方式，即：将数据集中的正类样本和负类样本分别随机分为 5 等份，两两随机组合得到 5 个大小一致的子集，将其中一份作为测试集，其余 4 个子集作为训练集，重复 5 次，以平均值作为最终的分类结果。

从表 3 可以看出，在少数类的识别评价度量 *F-measure* 值方面，本文算法具有明显优势：8 组 UCI 数据集中的 6 组精度最高，特别是在高度不平衡度的 nuersery、page-blocks 数据集上，与所比较的三种算法中的最优算法有 5% 以上的精度提升。与传统 AdaBoost 算法相比，在低不平衡度条件下，本文算法与之差异不明显，但随着不平衡度的增加，本文算法精度较高，例如在 letter、nursery、page-blocks 数据集上分别有 30.7%、19.8%、31.1% 的提升。另外，随着数据集不平衡度的增加，数据采样方法与 Boost 技术相结合的提升方法中，SMOTEBoost 算法的少数

表 3 各种方法的 *F-measure* 值比较

Table 3. *F-measure* values on test datasets

数据集	AdaBoost	SMOTEBoost	RUSBoost	本文算法
pima_diabetes	0.682	0.649	0.653	0.71
waveform	0.807	0.816	0.819	0.854
splice	0.932	0.946	0.925	0.938
artificial	0.593	0.652	0.641	0.66
optdigits	0.651	0.825	0.774	0.819
letter	0.69	0.885	0.72	0.902
nursery	0.726	0.828	0.78	0.87
page-blocks	0.53	0.654	0.55	0.695

表 4 各种方法的 *G-mean* 值比较

Table 4. *G-mean* values on test datasets

数据集	AdaBoost	SMOTEBoost	RUSBoost	本文算法
pima_diabetes	0.45	0.528	0.527	0.603
waveform	0.735	0.758	0.772	0.826
splice	0.935	0.934	0.933	0.93
artificial	0.525	0.66	0.771	0.818
optdigits	0.869	0.93	0.858	0.932
letter	0.849	0.967	0.86	0.94
nursery	0.892	0.905	0.95	0.87
page-blocks	0.447	0.772	0.85	0.81

类识别性能比 RUSBoost 算法更好。

表 4 给出了体现分类器对不平衡数据集的整体分类效果的评价。从中可以看出，在低度不平衡度和中度不平衡度条件下，本文算法的 *G-mean* 值在整体上最优；而在高度不平衡条件下，*G-mean* 值度量指标稍逊于所比较的算法，主要原因是在高度不平衡范围下，所比较的三类算法中的分类器倾向于忽略正类样本，在降低了体现少数类识别准确率的 *F-measure* 值情况下，提高了整体分类性能的 *G-mean* 值。

5　结束语

本文提出一种在无监督聚类基础上的改进 AdaBoost 算法用于处理不平衡数据分类。该方法首先进行基于聚类的欠采样处理，对初始训练集上的负类样本进行无监督的 K 均值聚类；同时借鉴代价敏感学习思想，对 AdaBoost 算法进行了改进，对基分类器的不同类别样本分类误分赋予不对称代价，在损失一定程度多数类分类性能的情况下，提高少数类的分类精度，以更符合实际的应用情况。实验结果表明，该方法在显著降低实际参与模型训练样本数量的同时，能够取得不错的分类性能，为大规模不平衡数据集分类问题提供了一种新的方法。

由于数据集本身的多样性和复杂性，样本的分布也呈现多样性，如果能实现估计正负类样本潜在的分布，根据不同的潜在分布设置不同的聚类方式，对算法的分类性能将会提高更多。

参 考 文 献

[1] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[2] Tomek I. Two modifications of CNN [J]. IEEE Transactions on Systems, Man and Communications, 1976, 6(11): 769-772.

[3] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[4] Yen SJ, Lee YS. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset [C] // International Conference on Intelligent Computing, Lecture Notes in Control and Information Sciences, 2006: 731-740.

[5] 蒋盛益, 苗邦, 余雯. 基于一趟聚类的不平衡数据下抽样算法 [J]. 小型微型计算机系统, 2012, 33(2): 232-236.

[6] He HB, Garcia EA. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[7] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述 [J]. 计算机科学, 2010, 37(10): 27-32.

[8] 李雄飞, 李军, 董元方, 等. 一种新的不平衡数据学习算法 PCBoost [J]. 计算机学报, 2012, 35(2): 202-209.

[9] 凌晓峰, Sheng VS. 代价敏感分类器的比较研究 [J]. 计算机学报, 2007, 30(8): 1203-1212.

[10] Fan W, Stolfo S, Zhang J, et al. AdaCost: misclassification cost-sensitive boosting [C] // Proceedings of the 16th International Conference on Machine Learning, 1999: 97-105.

[11] Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: comparison and improvements [C] // Proceedings of the 1st IEEE International Conference on Data Mining, 2001: 257-264.

[12] Chawla NV, Lazarevic A, Hall LO, et al. SMOTEBoost: improving prediction of the minority class in boosting [C] // Proceedings of the 7th European Conference Principles and Practice of Knowledge Discovery in Databases, 2003: 107-119.

[13] 林智勇, 郝志峰, 杨晓伟. 若干评价准则对不平衡数据学习的影响 [J]. 华南理工大学学报(自然科学版), 2010, 4(38): 126-135.