一种基于聚类融合欠抽样的不平衡数据分类方法

张枭山 罗 强
(重庆邮电大学计算机科学与技术学院 重庆 400065) (重庆邮电大学移通学院计算机系 重庆 401520)

摘 要 在面对现实中广泛存在的不平衡数据分类问题时,大多数传统分类算法假定数据集类分布是平衡的,分类结果偏向多数类,效果不理想。为此,提出了一种基于聚类融合欠抽样的改进 AdaBoost 分类算法。该算法首先进行聚类融合,根据样本权值从每个簇中抽取一定比例的多数类和全部的少数类组成平衡数据集。使用 AdaBoost 算法框架,对多数类和少数类的错分类给予不同的权重调整,选择性地集成分类效果较好的几个基分类器。实验结果表明,该算法在处理不平衡数据分类上具有一定的优势。

关键词 机器学习,不平衡数据,聚类融合,欠抽样,集成学习
中图法分类号 TP309 **文献标识码** A

Unbalanced Data Classification Algorithm Based on Clustering Ensemble Under-sampling

ZHANG Xiao-shan LUO Qiang
(Institute of Computer Science & Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)
(Department of Computer, College of Mobile Telecommunications, Chongqing University of Posts and Telecommunications, Chongqing 401520, China)

Abstract Imbalanced data exists widely in the real world, under such circumstances, most traditional classification algorithms assume the balanced data distribution, which results in the classification outcome offset to the majority class, so the effort is not ideal. The enhanced AdaBoost based on the clustering ensemble under-sampling technique was proposed in this paper. The algorithm firstly clusters the sample data by clustering ensemble, according to the sample weight. And the majority class from each cluster in certain proportion are randomly selected and then merge with all minority class to generate a balanced training set. By use of the AdaBoost algorithm framework, the algorithm gives different weight adjustment to the majority class and the minority class respectively, and selectes several base classifiers with better effect to get the final ensemble. The experiment result show that: this algorithm has a certain advantage dealing with unbalanced data classification.

Keywords Machine learning, Imbalanced data, Clustering ensemble, Under-sampling, Ensemble learning

1 引言

作为机器学习领域非常重要的一个研究方向,分类问题得到了学者的广泛研究。大多数传统分类算法通常假设数据集类分布是平衡的,将样本总体正确率作为其评价标准。然而在现实世界里,数据的类分布往往是不平衡的,如医疗诊断[1]、信用卡非法交易监测[2]、从卫星雷达图像上检测石油泄漏[3]等,在这些领域,我们真正关心的是少数类而不是多数类,因为其错分类代价更大,使得其识别正确率更为重要。

现今,对不平衡数据分类研究主要集中在以下几个方面。

(1)数据层面的处理,即如何将不平衡数据转化为平衡数据,如过抽样和欠抽样技术。在过抽样研究中,由 Chawla[4] 提出的 SMOTE 最为经典,现有的很多过抽样算法都以此为基础。SMOTE 是寻找样本的近邻集,从近邻集中随机选择样本与之形成连线,并在线上合成新样本点,然而,SMOTE 在合成样本的过程中具有一定的盲目性,它不能对合成样本数量进行精确控制,同时也没有充分考虑到多数类近邻样本,往往导致严重的样本重叠。针对 SMOTE 的不足,Han 等人[5]提出了 Borderline-SMOTE 算法,其基本思想是只对少数类的边界样本点进行 SMOTE 过抽样处理,没有对多数类进行处理,有效避免了冗余样本的产生。但如何有效判断少数类的边界有待进一步解决。刘余霞等人[6]提出 DB_SMOTE,其算法依据是假设位于类边缘的样本更有助于形成分类边界,通过直接比较样本与类中心距离和聚集程度得到种子样本,并在种子样本和类中心的连线上合成新样本,实现过抽样策略。Kubat 等人[7]的研究表明,欠抽样比过抽样的效果更好。程险峰等人[8]通过保留多数类边界附近样本,实现欠抽样策略,其难点在于领域半径不好确定。Yen 等人[9]提出的聚类欠抽样 SBC 系列算法获得了比较好的效果。

(2)算法层面的改进,修改已有的算法或设计新的算法,如代价敏感学习、集成学习等。集成分类器往往比单一分类器具有更高的准确率,常见的集成方法有袋装(Bagging)、提

张枭山(1990—)男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:zhangxiaoshanha@163.com;罗 强(1987—),男,助教,主要研究方向为智能缩微车、机器学习、云模型,E-mail:502842134@qq.com。

升(Boosting)和随机森林(Random Forests)等。其中非常有名的是 Freund 和 Schapire 提出的 AdaBoost[10] 法，SUN 等人[11]将代价参数引入到 AdaBoost 法的权值更新公式中，通过提高少数类样本代价，使得后续生成的基分类器更加关注少数类，提高了分类精度。

(3)数据层面和算法层面同时调整，Seiffert 等人[12]提出的 RUSBoost 算法运用随机欠抽样从多数类中抽取样本，将获得的多个子分类器进行集成，算法简单易实现，且可以显著提高分类系统的泛化能力[13]。结合 SMOTE 过抽样方法，Chawla 等[14]提出 SMOTEBoost 算法，它将合成样本引入到每一次的迭代过程中，保证各个基分类器更多地获取少数类样本信息，提高了集成分类器的性能。李雄飞等[15]提出的 PCBoost 算法，在每次迭代初始，利用随机采样的数据合成方法，添加合成的少数类样本，平衡训练信息，并及时进行“扰动修正”，删除错分的合成样本，在处理不平衡数据分类问题上具有一定的优势。

聚类融合[16-19]是一种较新的聚类方法，它通过合并数据集的多次聚类结果获得一个更优的聚类划分。与单一聚类算法相比，聚类融合算法具有更好的鲁棒性、适用性、稳定性、并行性和可扩展性[20]。在聚类融合中，先要产生样本集合 D 的 H 个聚类结果，然后将 H 个聚类成员的聚类结果合并。一般使用 MacQueen[21] 提出的 K 均值算法来产生聚类成员，因为 K 均值算法的计算复杂度低，而且对初始簇中心的选择很敏感。

本文提出了一种以聚类融合欠抽样和 Boosting 技术相结合的不平衡数据分类方法——基于聚类融合欠抽样改进的 AdaBoost 算法。算法首先进行聚类融合，在每轮迭代的时候，对每个聚类簇中的多数类进行欠抽样，从中抽取一定比例的多数类和所有的少数类组成平衡的训练集用于训练弱分类器，样本被抽中的概率与其权值正相关，借鉴代价敏感的思想，对错分类的少数类样本给予更大的权值，在获得的分类器中选择性地集成，从而得到最终的分类模型，模型的泛化能力可得到显著提高。

2 相关知识

2.1 聚类算法

聚类分析是一种无监督学习方法。聚类是把未标记的数据对象按照其相似度划分为多个有意义的簇的过程，使得簇内部数据对象具有高相似度，而不同簇中对象具有较小的相似度。

K 均值是一种聚类算法，它以样本相似度为标准，迭代地更新聚类的簇中心，当簇中心趋于收敛时，则停止迭代，获得最终的聚类结果。

K 均值算法的步骤如下：

(1)随机地从样本集中选择 k 个对象作为初始聚类中心。

(2)根据度量样本相似度的公式，将数据集中其他的样本分配到离它最近的簇中。

(3)重新计算各个簇类中心。

(4)如果簇类中心不再变化，则算法结束；否则转第(2)步。

K 均值算法由于对初始簇中心的选择很敏感，单纯地使用此方法来聚类，聚类结果不稳定，鲁棒性也不是很好。

2.2 重抽样

重抽样是通过改变训练样本的分布来消除或减小训练样本集的不平衡度，包括欠抽样和过抽样。欠抽样是从多数类样本中抽取部分参与训练，单纯地使用此方法可能造成样本信息丢失；而过抽样，则是按照一定的策略合成出一部分少数类样本，同样，单独地使用此方法，算法很容易出现过学习。一个好的重抽样方法，其抽选的样本要具有代表性，能较好地反映初始样本的分布特征。

2.3 AdaBoost 算法

AdaBoost 算法是一种典型的集成学习方法，可以有效地提高学习模型的泛化能力。它首先赋予每个训练样本以相同权值，算法迭代若干轮得到若干弱分类器；对于训练错误的样本，算法增加其权值，也就是让后续弱分类器更关注这类“困难”样本，对于训练正确的样本，算法减小其权值，以降低下一轮被弱分类器选中的机会；最后通过对这些弱分类器加权求和集成最终的分类器。

AdaBoost 算法描述如下：

给定训练集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 和预定的迭代次数 T，$w^t(i)$ 表示第 t 轮迭代中样本 x_i 的权值。

Step 1　初始化样本权值：

$$w^1(i)=\frac{1}{N},i=1,2,\cdots,N \tag{1}$$

Step 2　For $t=1,2,3,\cdots,T$

Step 2.1　使用弱分类算法在带权样本上训练得到弱分类器 h_t。

Step 2.2　计算 h_t 在当前样本分布上的加权错误率：

$$\varepsilon_t=\sum_{i=1}^{n}w^t(i),y_i\neq h_t(x_i) \tag{2}$$

Step 2.3　如果 $\varepsilon_t>0.5$

$$t=t-1 \tag{3}$$

转 Step 2。

Step 2.4　计算弱分类器 h_t 的权值：

$$\partial_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{4}$$

Step 2.5　更新训练样本权值：

$$w^{t+1}(i)=\frac{w^t(i)\exp(-\partial_t y_i h_t(x_i))}{Z_t} \tag{5}$$

其中，Z_t 为归一化因子，使得 $\sum_{i=1}^{N}w(i+1)=1$。

Step 3　输出最后强分类器：

$$H(x)=sign\left(\sum_{i=1}^{N}\partial_i h_i(x)\right) \tag{6}$$

在处理不平衡数据集时，由于训练样本类分布的不平衡性，且算法将多数类与少数类同等对待，分类器的分类结果偏向于多数类，算法对少数类分类的精度不高。

3 基于聚类融合欠抽样改进 AdaBoost 算法

为便于讨论，本文主要关注二类非平衡数据分类问题。在此情况下，通常称多数类为负类，少数类为正类，对应的类别标签取值分别为{−1，+1}。

算法包括以下几个阶段：初始化训练样本权值；使用聚类

融合算法对训练样本聚类；算法迭代一定的轮数，根据簇中多数类与少数类的比率与多数类的权值，按照一定比例从每个簇中抽取负类与全部的正类合成训练样本集；分类错误时，对正类样本和负类样本给予不同的权值修正；从获得的分类器中筛选出差异度大的选择性集成，输出最后的分类器。

给定样本集合 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中，$n$ 为训练集样本总数，x_i 是输入空间 X 的实例，$y_i\in\{-1,+1\}$ 是输出分类 Y 对应的分类标签，迭代次数为 T。算法流程如下：

Step 1　重复使用 K 均值算法产生 h 个聚类结果，共识函数使用 Fred[22] 提出的 Co-association 矩阵方法，最终将训练样本聚成 c 个簇，分别记为 $C_1,C_2,\cdots,C_c$。

Step 2　按照式(1)初始化样本权重。

Step 3　For $t=1,2,3,\cdots,T$

Step 3.1　根据各个簇中负类数 $MajSize_{C_i}$ 与正类 $MinSize_{C_i}$ 的比率，从每个簇中抽取部分负类 $ExtraMajSize_{C_i}$ 与所有正类合并成 1∶1 平衡数据集用于下面的分类器训练。每个簇中负类样本被抽中的概率与样本权重相关，$MajSize$ 为负类样本总数，则每个簇中抽取的负类样本数：

$$ExtraMajSize_{C_i}=MajSize\frac{MajSize_{C_i}/MinSize_{C_i}}{\sum_{i=1}^{c}MajSize_{C_i}/MinSize_{C_i}} \tag{7}$$

Step 3.2　使用弱学习算法训练得到分类器 h_t。

Step 3.3　按照式(2)计算 h_t 在当前样本分布上的训练误差。

Step 3.4　按照式(4)计算弱分类器 h_t 的权值。

Step 3.5　样本被正确分类：

$$w^{t+1}(i)=\frac{e^{-\partial_t}w^t(i)}{Z_t} \tag{8}$$

负类样本被错误分类：

$$w_-^{t+1}(i)=\frac{e^{\partial_t}w^t(i)}{Z_t} \tag{9}$$

正类样本被错误分类：

$$w_-^{t+1}(i)=\frac{\sqrt{2}e^{\partial_t}w^t(i)}{Z_t} \tag{10}$$

Step 4　通过遗传算法选择出差异度比较大的分类器，选择性集成得到最后的强分类器：

$$H(x)=sign(\sum_{i=1}^{NUM}\partial_i h_i(x))$$

其中，NUM 为最后集成的分类器数目。

该算法首先聚类融合，相比单一聚类算法，聚类效果更好、更稳定；再使用欠抽样平衡训练集的样本分布，相比随机欠抽样，基于聚类的欠抽样可以更好地抽取具有代表性的样本，抽取的样本能较好地反映原始样本分布；然后，模型借鉴 AdaBoost 算法框架，算法根据分类器对正类和负类不同的分类情况，调整其权值，并使得错分类的正类具有更大的权重，以便让后续训练过程更加关注这类正类样本的分类；最后使用选择性集成策略，得到最后的强分类器。

4　实验与分析

4.1　数据集

为了验证算法的有效性，选取了 6 组 UCI 数据集，每个数据集样本总数、负类样本数、正类样本数及不平衡度如表 1 所列。对于含有多个类别的数据集，将其中一类作为正类，其余类统一作为负类。由于其提供者要求，Satimage 数据集不进行十折交叉验证，其余均采用十折交叉验证。

表 1　UCI 数据集

数据集	样本总数	正类	负类	不平衡度
Glass	214	29	185	1.16
Pima	768	268	500	1.87
Haberman	306	81	225	2.78
Vehicle	846	212	634	2.99
Satimage	6435	626	5809	9.28
Letter	20000	734	19266	26.24

4.2　不平衡数据分类的评价方法

传统分类算法一般使用分类精度(即正确分类的样本数占到总样本个数的比率)作为评价指标。然而，对于不平衡数据，以分类精度作为评估分类器性能是不合理的，极端来说，如果 1%的信用卡交易是欺诈行为，则预测每个交易都合法的模型具有 99%的准确率，然而，这样的预测是没有实际意义的，它检测不出任何的欺诈交易，对于这类问题，我们往往更加关心对正类样本的分类。因此，对于不平衡数据分类问题，需要一个合理的评价指标来验证分类器的有效性。

现今，不平衡数据评价方法大多基于混淆矩阵(见表 2)，则 TP 和 TN 分别表示被分类器正确分类的正类数和被分类器正确分类的负类数，而 FN 和 FP 分别表示被分类器错误分类的负类数和被分类器错误分类的正类数。

表 2　二分类混淆矩阵

	预测为正类	预测为负类
正类样本	TP	FN
负类样本	FP	TN

相关评价指标：准确率 precision、召回率 recall、G-means 及 F-measure 计算如下：

$$precision=\frac{TP}{TP+FP} \tag{11}$$

$$recall=\frac{TP}{TP+FN} \tag{12}$$

$$G\text{-}means=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{13}$$

$$F\text{-}measure=\frac{(\beta^2+1)recall\cdot precision}{\beta^2\cdot recall+precision} \tag{14}$$

准确率又称为正确率，表示正类样本被正确分类的比率，对样本的分布敏感。召回率又称全查率，表示正类样本被正确分类的完整度，用来衡量分类器对正类样本的敏感程度。理想情况下，准确率和召回率越高，分类效果越好，但在现实中，这两个指标往往是矛盾的。

当参数 $\beta=1$ 时，就是最常见的 F1 值，F1 值是准确率和召回率的加权调和平均，F1 高时表示分类结果比较理想。G-means 是取正类分类准确率和负类分类准确率的均衡值。

本文使用 F1 和 G-means 作为评价指标，其中 F1 用来衡量对正类正确分类的能力，G-means 用来衡量整体分类性能。

4.3　实验结果

本文的仿真实验基于 Weka 平台，在 Eclipse 环境中实现，并且与传统分类算法 SMOTE、AdaBoost、SMOTEBoost

和 RUSBoost 进行了比较，各个算法对应的 F1 值和 G-均值相关的对比结果如表 3 和表 4 所列。实验中，上述对比算法使用的弱分类器均为 J48，本文提出的算法则采用在小样本平衡数据集上表现优异的支持向量机(Support Vector Machine)系列算法中的 LIBSVM 作为弱分类器。

表 3　各种方法的 F1 值

数据集	SMOTE	AdaBoost	SMOTEBoost	RUSBoost	本文
Glass	0.755	0.813	0.846	0.892	0.969
Pima	0.634	0.681	0.648	0.653	0.711
Haberman	0.421	0.369	0.391	0.385	0.523
Vehicle	0.614	0.554	0.617	0.621	0.596
Satimage	0.585	0.647	0.678	0.686	0.631
Letter	0.849	0.691	0.874	0.725	0.901

表 4　各种方法的 G-means

数据集	SMOTE	AdaBoost	SMOTEBoost	RUSBoost	本文
Glass	0.786	0.894	0.911	0.917	0.949
Pima	0.711	0.703	0.724	0.718	0.736
Haberman	0.574	0.523	0.556	0.541	0.622
Vehicle	0.744	0.668	0.734	0.752	0.684
Satimage	0.863	0.756	0.761	0.797	0.745
Letter	0.955	0.848	0.966	0.863	0.942

可以看出，在较低不平衡度数据集 Glass、Pima、Haberman 上，本算法的两个度量指标均优于其他算法；而在高度不平衡的数据集 Letter 上，一个度量指标均优于其他算法，并且另外一个指标也有不俗的表现。实验结果表明：本算法在处理不平衡数据分类上是有效的。

结束语　本文提出了一种新的不平衡数据分类算法，算法将聚类融合欠抽样与 Boosting 技术结合起来，将不平衡数据分类问题转换为对多个平衡数据子集的分类问题；借鉴代价敏感方法，使用 AdaBoost 算法框架对基分类器的不同类别样本错分类赋予不同代价；最后使用选择性集成策略得到最终的分类器，提高了对正类的分类精度。实验结果表明，与传统算法相比，本文提出的算法在处理不平衡数据分类问题上具有一定的优势。如何更好地得到聚类结果、加快算法的收敛速度是今后需要进一步研究的内容。

参 考 文 献

[1] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[2] Chan P K, Stolfo S J. Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection[C]// KDD. 1998: 164-168

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

[4] 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(1): 321-357

[5] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]// Advances in intelligent computing. Springer Berlin Heidelberg, 2005: 878-887

[6] 刘余霞，刘三民，刘涛，等. 一种新的过采样算法 DB_SMOTE [J]. 计算机工程与应用，2014，50(6)：92-95

[7] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection[C]// ICML. 1997: 179-186

[8] 程险峰，李军，李雄飞. 一种基于欠采样的不平衡数据分类算法 [J]. 计算机工程，2011，37(13)：147-149

[9] Yen S J, Lee Y S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718-5727

[10] 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

[11] Sun Y, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378

[12] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. RUSBoost: improving classification performance when training data is skewed[C]// 19th International Conference on Pattern Recognition, 2008(ICPR 2008). IEEE, 2008: 1-4

[13] Ditterrich T G. Machine learning research: four current direction [J]. Artificial Intelligence Magzine, 1997, 18(4): 97-136

[14] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[M]// Knowledge Discovery in Databases (PKDD 2003). Springer Berlin Heidelberg, 2003: 107-119

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

[16] Minaei-Bidgoli B, Topchy A P, Punch W F. A Comparison of Resampling Methods for Clustering Ensembles[C]// IC-AI. 2004: 939-945

[17] Hadjitodorov S T, Kuncheva L I, Todorova L P. Moderate diversity for better cluster ensembles[J]. Information Fusion, 2006, 7(3): 264-275

[18] Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J]. The Journal of Machine Learning Research, 2003, 3: 583-617

[19] Fred A L N, Jain A K. Data clustering using evidence accumulation[C]// Proceedings 16th International Conference on Pattern Recognition, 2002. IEEE, 2002: 276-280

[20] Topchy A, Jain A K, Punch W. A mixture model of clustering ensembles[C]// Proc. SIAM Intl. Conf. on Data Mining. 2004

[21] MacQueen J. Some methods for classification and analysis of multivariate observations[J]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(14): 281-297

[22] Fred A. Finding consistent clusters in data partitions[M]// Multiple classifier systems. Springer Berlin Heidelberg, 2001: 309-318