剪枝与欠采样相结合的不平衡数据分类方法 *

张 健, 方宏彬
(安徽大学 数学科学学院, 合肥 230601)

摘 要: 通过剪枝技术与欠采样技术相结合来选择合适数据,以提高少数类分类精度,研究欠采样技术在不平衡数据集环境下的影响。结果表明,与直接欠采样算法相比,本文算法不仅在 accuracy 值上有所提高,更重要的是大大改善了 g-means 值,特别是对非平衡率较大的数据集效果会更好。

关键词: 机器学习; 不平衡数据集; 剪枝技术; 欠采样技术; 交叉验证; 合并分类器增强算法
中图分类号: TP311.13 **文献标志码**: A **文章编号**: 1001-3695(2012)03-0847-02
doi:10.3969/j.issn.1001-3695.2012.03.012

Pruning and undersampling combination of imbalanced data classification method

ZHANG Jian, FANG Hong-bin
(*School of Mathematical Sciences*, *Anhui University*, *Hefei* 230601, *China*)

Abstract: This paper proposed pruning and under-sampling combined approaches for selected the representative data as training data to improve the classification accuracy for minority class and investigated the effect of under-sampling methods in the imbalanced class distribution environment. The experimental results show that the accuracy of algorithm of this paper compare with direct undersampling algorithm have increased, the most important is to significantly improve the g-means value. Especially, the effect will be better on the imbalance rate of larger data sets.

Key words: machine learning; imbalanced data sets; pruning techniques; under-sampling; cross-validation; AdaBoost algorithm

分类问题是数据挖掘、机器学习和模式识别中十分重要的研究课题之一,不平衡数据集分类问题又是该课题的热点问题。不平衡数据集是指数据集中某些类的样本比其他类多很多,样本多的类为多数类(即负类),样本少的类为少数类(即正类)[1]。目前国内外已取得大量的成果,其中采样技术应用较广泛。早在 1997 年,Kubat 等人[2]就选择对多数类数据进行欠采样,以达到与少数类数据平衡,Yen 等人[3,4]也是采用欠采样技术。另一方面就是过采样技术,Japkowicz[5,6]采用了 under-sampling、resampling 和 arecognition-based induction scheme 三种策略,其主要目的都是降低数据集的不平衡程度。但是欠采样容易使重要信息丢失,而过采样会重复复制。Chawla 等人[7]改进过采样方法,提出了 SMOTE(synthetic minority oversampling technique),继而又提出了 SMOTEBoost 方法[8],不仅克服了过采样的重复复制,还提高了分类器的精确度。

欠采样技术虽然降低了数据集的不平衡程度,但是在实际应用中极容易使多数类数据重要信息数据丢失。为此,本文提出了剪枝技术和欠采样技术相结合的方法。

1 剪枝技术

园艺工人通过修剪多余的枝叶培植出各种造型的盆景树木,让人赏心悦目。剪枝技术思想就源于园艺工人修剪树木。首先来看下面两张图,图 1 是没经过剪枝的,图 2 是经过剪枝的,图 2 显然容易被分类器学习。那么剪枝技术是怎样剪枝的呢?

可以认为不平衡数据集来源于两类不同分布机制,一类产生多数类(MA),另一类产生少数类(MI)。本来两种分布机制可以产生同样多的 MA 和 MI,但是现实中 MA 分布机制产生的数据要远多于 MI,这样就不便于分类器学习,所以进行适当剪枝。把 MA 分成三类:a)绝对安全数据,该类数据来源于 MA 分布机制产生的概率远大于 MI 分布机制产生的概率;b)边缘数据,该类数据来源于 MA 和 MI 分布机制产生的概率相等;c)噪声数据,该类数据来源于 MA 分布机制产生的概率远小于 MI 分布机制产生的概率。剪枝就是剪去噪声数据。剪枝具体算法如下:

输入:所有数据集,多数类(MA)和少数类(MI),多数类数据个数 N,不平衡率 $k(k=\mathrm{MA}/\mathrm{MI})$。
输出:分成三类的多数类(MA),少数类(MI)。
执行循环:for $i=1,\cdots,N$
找出 i 的 k 个近邻数据
如果这 k 个数据都是多数类,则 i 是绝对安全数据
如果这 k 个数据有一个是少数类,则 i 是边缘数据
其他情况,则 i 是噪声数据

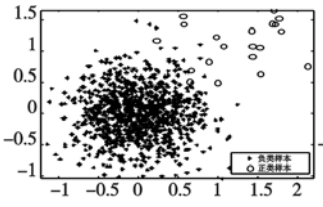

图1 未经过剪枝

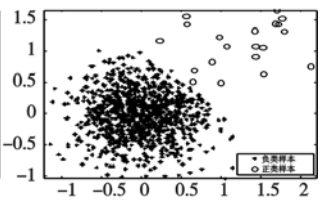

图2 经过剪枝

2 欠采样技术

欠采样技术在不平衡数据集分类中的应用目的是压缩多数类样本来降低不平衡程度,但并不是仅仅去掉部分多数类样

收稿日期: 2011-08-31; **修回日期**: 2011-10-17 **基金项目**: 国家自然科学基金资助项目(71071002);安徽省教育厅自然科学基金资助项目(05010428);安徽大学人才队伍建设项目;安徽大学学术创新团队项目(KJTD001B)

作者简介:张健(1981-),男,安徽安庆人,硕士研究生,主要研究方向为机器学习、模式识别(zj520zj@163.com);方宏彬(1972-),男,安徽池州人,副教授,博士,主要研究方向为智能计算、信息融合.

本,而是选择足够的富含重要信息的样本。多数类数据通过剪枝后,留下绝对安全数据和边缘数据,仅对绝对安全数据进行欠采样。欠采样技术是根据张铃提取的商空间理论进行网格划分[9,10]。具体算法如下:

输入:需要采样的样本。
输出:欠采样后的样本。
执行:对多数类(MA)所在空间进行网格划分,生成 n 个网格,有 m 个网格内有样本(一般 $n > m$)。
执行循环 for $i = 1, \cdots, m$
求 i 网格中样本均值 mean
如果 i 网格只需取一个样本,那就是样本均值 mean
如果 i 网格不止一个,比方 k 个,就取样本均值 mean 的 k 个近邻样本

3 剪枝与欠采样相结合

本文通过剪枝技术修剪多数类样本得到三类数据:绝对安全数据、边缘数据和噪声数据。去掉噪声数据,留下绝对安全数据和边缘数据,再通过上文中欠采样技术对安全数据进行压缩,降低不平衡度,最后把得到的多数类和少数类样本合并进行 AdaBoost 学习[11]。根据上述思想,本文的算法步骤如下:a)用剪枝技术修剪多数类样本得到绝对安全数据、边缘数据和噪声数据三类;b)用欠采样技术对安全数据进行压缩,降低不平衡度;c)再用 AdaBoost 学习,得到最终分类器。

4 实验分析

4.1 数据集说明及评价标准

为了验证剪枝与欠采样相结合方法的效率,在典型的 UCI 数据库集进行测试,分别采用了 Pima、blood-transfusion 和 satimage 数据集。测试中把 satimage 数据集转换成两类,取类标为"4"的一类为少数类,其他合并一类为多数类,如表 1 所示。

表 1 实验采用的数据集

数据集	属性数	总样本数	构成
pima	8	768(268/500)	0:1
blood-transfusion	4	748(178/570)	0:1
satimage	36	4435(415/4020)	4: other

4.2 评价标准

一般情况下,正确分类率能准确反映学习器的泛化性能,其定义为

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

各符号含义如表 2 所示。

表 2 相关符号含义

样本	实际正类样本	实际负类样本
预测的正类样本	TP(正确的正类)	FN(错误的负类)
预测的负类样本	FP(错误的正类)	TN(正确的负类)

对于不平衡数据,这一指标实际意义并不大,因为它反映的是多数类样本的分类测试结果,所以本文还是采用 g-means 指标。g-means 值是不平衡数据集常采用的衡量指标,它可以有效地衡量不平衡数据的分类精度,一般 g-means 值越大分类效果越好,其定义为

$$\text{g-means} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{2}$$

4.3 实验验证及分析

采用交叉验证(cross-validation)的方法,把所有样本随机分成五份,并且保持每份样本与原来样本平衡率大致一样。每次取四份作为训练集,剩下的一份作为测试集,计算测试集的各项评价指标值,然后把五次值的平均值作为本算法的评价指标值。使用直接欠采样算法与本文算法进行比较,图 3 ~ 5 分别是三类数据集的 accuracy 值,图 6 ~ 8 分别是三类数据集的 g-means 值。可以看出,本文算法在不同划分情况下对于数据集 Pima 和 blood-transfusion 的 accuracy 值和 g-means 值明显优越于直接欠采样算法。虽然 satimage 数据集的 accuracy 值比直接欠采样算法要低一点,但是 g-means 值明显优越于直接欠采样算法。

5 结束语

本文提出处理不平衡数据的剪枝与欠采样相结合分类算法,对多数类数据进行修剪再进行欠采样,平衡数据集。结果表明,其在一定程度上提高了分类器对少数类的分类精度,为不平衡数据集分类问题提供一种新的研究思路。

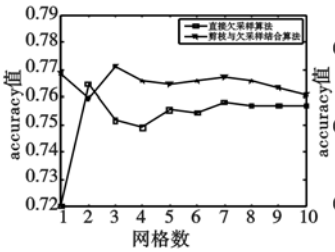

图3 数据Pima两种算法accuracy值比较

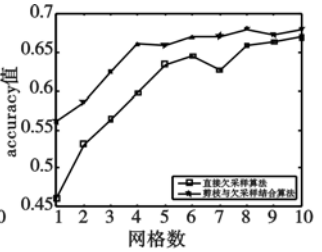

图4 数据blood-transfusion两种算法accuracy值比较

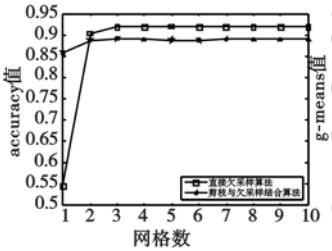

图5 数据satimage两种算法accuracy值比较

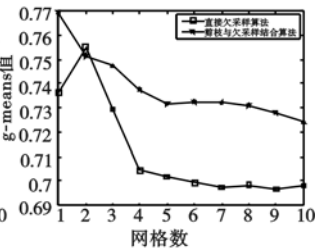

图6 数据pima两种算法g_means值比较

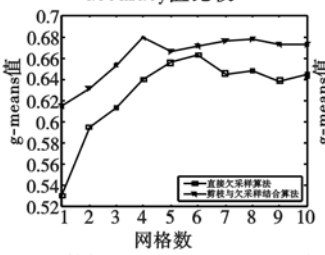

图7 数据blood-transfusion两种算法g-means值比较

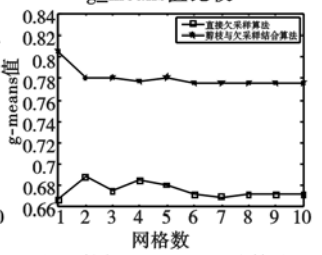

图8 数据satimage两种算法g-means值比较

参考文献:

[1] WEISS G M. Mining with rarity:a unifying framework[J]. SIGKDD Explorations,2004,6(1):7-19.
[2] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one sided selection[C]//Proc of the 14th International Conference on Machine Learning. 1997:179-186.
[3] 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.
[4] 郭虎,升亓慧,王文剑. 处理非平衡数据的粒度 SVM 学习算法[J]. 计算机工程, 2010,36(2):181-183.
[5] JAPKOWICZ N. The class imbalance problem: significance and strategies[C]//Proc of International Conference on Artificial Intelligence. 2000.
[6] JAPKOWICZ N. Concept-learning in the presence of between-class and within class imbalances[C]//Proc of the 14th Conference of the Canadian Society for Computational Studies of Intelligence. 2001:67-77.
[7] 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.
[8] CHAWLA N V,LAZAREVIC A,HALL O. SMOTEBoost: improving prediction of the minority class in boosting[C]//Proc of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin:Springer,2003:107-119.
[9] 张钹,张铃. 问题求解理论及应用——商空间粒度计算理论及应用[M]. 2 版. 北京: 清华大学出版社,2007.
[10] 文贵华,向君,丁月华. 基于商空间粒度理论的大规模 SVM 分类算法[J]. 计算机应用研究, 2008,25(8):2299-2301.
[11] FREUND Y, SCHAPIRE R. Experiments with a new boosting algorithm[C]//Proc of the 13th International Conference on Machine Learning. 1996:148-156.