文章编号:1001 - 9081(2006)08 - 1894 - 04

不均衡数据集学习中基于初分类的过抽样算法

韩 慧,王 路,温 明,王文渊
(清华大学 自动化系,北京 100084)
(hanh01@mails.tsinghua.edu.cn)

摘 要:为了有效地提高不均衡数据集中少数类的分类性能,提出了基于初分类的过抽样算法。首先,对测试集进行初分类,以尽可能多地保留多数类的有用信息;其次,对于被初分类预测为少数类的样本进行再次分类,以有效地提高少数类的分类性能。使用美国加州大学欧文分校的数据集将基于初分类的过抽样算法与合成少数类过抽样算法、欠抽样方法进行了实验比较。结果表明,基于初分类的过抽样算法的少数类与多数类的分类性能都优于其他两种算法。

关键词:不均衡数据集;过抽样;欠抽样
中图分类号: TP311.13 **文献标识码**:A

Over-sampling algorithm based on preliminary classification in imbalanced data sets learning

HAN Hui, WANG Lu, WEN Ming, WANG Wen-yuan
(*Department of Automation, Tsinghua University, Beijing* 100084, *China*)

Abstract: To significantly improve the classification performance of the minority class, an over-sampling algorithm based on preliminary classification was presented. Firstly, preliminary classification was made on the test data in order to save the useful information of the majority class as much as possible. Then the test data that were predicted to belong to minority class were reclassified to improve the classification performance of the minority class. Using the data sets provided by University of California, Irvine, the new algorithm was compared with synthetic minority over-sampling technique and under-sampling method. The experimental results show that the new algorithm performs better than the others in terms of the classification performance of the minority class and majority class.

Key words: imbalanced data sets; over-sampling; under-sampling

不均衡数据集的分类问题是机器学习和模式识别领域中新的研究热点,是对传统分类方法的重大挑战,解决这一问题可以提出新的分类学习思想,从而有利于完善机器学习体系。所谓不均衡数据集,是指同一个数据集中某些类的样本比其他类的样本多很多,其中样本多的类为多数类,样本少的类为少数类[1]。很多实际的应用领域中都存在不均衡数据集,例如信用卡欺诈检测、医疗诊断、信息检索和文本分类等。但是,对不均衡数据集进行分类时,传统的分类方法倾向于对多数类有较高的识别率,对少数类的识别率却很低。例如,在信用卡欺诈检测中,合法的信用卡用户(多数类)比欺诈信用卡用户(少数类)多得多,虽然将合法的信用卡误分类为欺诈信用卡,银行要投入额外的人力与物力来验证;但是,如果将欺诈信用卡误分类为合法信用卡,所带来的经济损失比上一种情况要大得多。因此,在上述应用领域中,人们更加关心的是不均衡数据集中的少数类样本,如何有效地提高少数类的分类性能成为机器学习和模式识别领域亟待解决的课题。

1 不均衡数据集分类问题的研究现状

1.1 评价准则

因为多类问题通常可以简化为两类问题来解决,所以不均衡数据集的分类问题的研究重点是提高两类问题中少数类的分类性能。表 1 是两类数据集的混合矩阵,它是机器学习与模式识别领域中评价分类性能的常用方法。第一列表示样本的真实类标号,少数类和多数类的真实类标号分别是 Positive 和 Negative;第一行表示分类器分类的类标号。TP 和 TN 分别表示正确分类的少数类和多数类的样本数量,FN 和 FP 分别表示误分类的少数类和多数类的样本数量。精确度(式(1))是分类问题中常用的评价准则,它反映分类器对于数据集的整体分类性能。但是,它不能正确地评价不均衡数据集的分类结果。这是因为,多数类样本比少数类样本多得多,如果将所有的样本都分类为多数类,精确度仍然很高,但是少数类的识别率为零。因此,我们需要更合理的评价准则。F-value(式(2))是不均衡数据集分类问题中有效的评价准则,它是查全率(Recall)和查准率(Precision)的组合,其中 β 是可以调节的,通常取值为 1。当查全率和查准率的值都大时,F-value 的值才会相应增大,所以 F-value 可以正确地评价分类器对于每一类的分类性能。此外,计算少数类的 F-value 时,$Recall = TP/(TP + FN)$,$Precision = TP/(TP + FP)$;计算多数类的 F-value 时,$Recall = TN/(TN + FP)$,$Precision = TN/(TN + FN)$。

收稿日期:2006 - 03 - 01;**修订日期**:2006 - 05 - 08

作者简介:韩慧(1976 -),女,山东新泰人,博士研究生,主要研究方向:机器学习与模式识别中不均衡数据集的分类; 王路(1980 -),男(回族),江苏扬州人,博士研究生,主要研究方向:数据挖掘; 温明(1979 -),男,山西太谷人,博士研究生,主要研究方向:机器学习与模式识别; 王文渊(1940 -),男,天津人,教授,博士生导师,主要研究方向:机器学习与模式识别.

表 1　两类数据集的混合矩阵

	Classified Positive	Classified Negative
Positive	TP	FN
Negative	FP	TN

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$F\text{-}value = \frac{(1+\beta^2)\cdot Recall \cdot Precision}{\beta^2 \cdot Recall \cdot Precision} \quad (2)$$

1.2　解决方法

为了提高少数类的分类性能,不均衡数据集分类问题的解决方法分为数据层的方法和算法层的方法。

数据层的方法对训练集进行预处理,之后,用处理过的数据集训练分类器。数据层的方法又称为重抽样,分为过抽样和欠抽样。过抽样方法通过增加少数类样本来提高少数类的分类性能,最简单的过抽样方法是复制少数类样本,缺点是没有给少数类增加任何新的信息,会使分类器学到的决策域变小,从而导致过学习;欠抽样方法通过减少多数类样本来提高少数类的分类性能,最简单的欠抽样方法是随机地去掉某些多数类样本来减小多数类的规模,缺点是会丢失多数类的一些重要信息。因此,人们提出了许多改进的重抽样方法。例如,在少数类中加入随机高斯噪声,或者产生新的合成少数类样本等方法可以在一定程度上避免随机过抽样中出现的过学习问题[2];去掉远离分类边界或者引起数据重叠的多数类样本,得到的分类效果会比随机欠抽样理想一些[3,4]。

算法层的方法对分类算法本身进行操作。对已有的分类算法,通过调节不同类样本之间的成本函数、改变概率密度、调整分类边界等措施使其更有利于少数类的分类[5~7];文献[8]介绍的学习算法只对少数类样本进行训练,其目标是从测试样本中识别出感兴趣的少数类样本,而不是对少数类和多数类进行区分。

2　基于初分类的过抽样算法

如上所述,不均衡数据集分类问题的解决方法中,数据层的方法是通过减轻数据集的不均衡程度来提高少数类的分类性能。但是,数据层的方法中,欠抽样方法与过抽样方法都有不足之处。首先,欠抽样方法随机或者有选择地丢掉训练集中的某些多数类样本,这样做导致的结果是:分类器没有学习到多数类的某些信息。因此,欠抽样方法虽然可以增加正确分类的少数类样本的数量(即 TP 增大,FN 减小),但是,同时也增加了误分类的多数类样本的数量(即 TN 减小,FP 增大)。也就是说,欠抽样方法虽然可以提高少数类的查全率($Recall = TP/(TP + FN)$),但是,由于两类样本数量相差较大,误分类的多数类样本的数量(FP)有时会比正确分类的少数类样本的数量(TP)还要大,这样会阻碍少数类的查准率($Precision = TP/(TP + FP)$)的提高,所以,欠抽样方法中少数类 F-value 的值不会得到很大的提高。其次,过抽样方法没有对多数类样本做任何处理,而是通过随机复制或者产生合成样本等方式来扩大少数类的规模,从而减轻数据集的不均衡程度。但是,当多数类与少数类的样本数量相差非常大时,少数类样本需要过抽样很多倍才能达到要求,这样不仅会增大计算量,而且会导致过学习。此外,算法层的解决方法通过调节分类算法的内部参数或者提出新的算法来提高少数类的分类精度,由此得到的效果等同于数据层的方法[1]。

基于以上分析,本文提出了基于初分类的过抽样算法 OSPC (Over-Sampling algorithm based on Preliminary Classification)。OSPC 与已有的过抽样和欠抽样方法不同,对不均衡数据集分类时,OSPC 既没有丢失多数类的有用信息,少数类的分类性能也得到了更大程度的提高。

假设已知训练集 $T = \{S, M\}$,其中 $S = \{s_1, s_2, \cdots, s_{numS}\}$ 为少数类的样本集合,$M = \{m_1, m_2, \cdots, m_{numM}\}$ 为多数类的样本集合。其中,$numS$ 和 $numM$ 为少数类和多数类的样本个数,并且 $numS \ll numM$。

OSPC 算法的预处理过程如图 1 所示。

图 1　OSPC 算法的预处理过程

1) 对于多数类样本($i = 1,2,\cdots,numM$),若它在训练集 T 中的 K 个近邻也都属于多数类,根据 k-近邻的思想[9],则 m_i 离分类边界较远,对分类是相对安全的。将所有满足上述条件的多数类样本放入集合 E,称 E 为多数类的“分类种子”集合。

2) 将少数类 S 和多数类的“分类种子”集合 E 合并为第一个新的训练集 $T\_1 = \{S, E\}$。

3) 利用 T_1 对多数类 M 的样本进行最近邻(1NN)分类,误分类的多数类样本放入集合 H,可以看出,$H \subset M$。将少数类 S 和集合 H 合并为第二个新的训练集 $T\_2 = \{S, H\}$。

OSPC 算法的分类过程如图 2 所示。

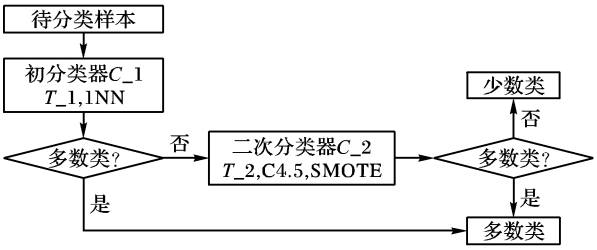

图 2　OSPC 算法的分类过程

1) 利用第一个新的训练集 T_1 作为最近邻算法的训练集,产生初分类器 C_1。

2) 利用合成少数类过抽样算法 SMOTE[2] 对 T_2 中的少数类 S 进行扩充,做法是:对于每个 $s_i(i = 1,2,\cdots,numS)$,找出它在 S 中的同类 k 近邻。之后,根据过抽样倍数的要求,从中任选 $n(n = 1,2,\cdots,k)$ 个近邻,按照式(3)产生新的合成少数类样本:

$$synthetic_j = s_i + r_j \times dif_j \qquad j = 1,2,\cdots n \quad (3)$$

其中 dif_j 是 s_i 和第 j 个近邻的矢量差,r_j 是 0 到 1 之间的随机数。通常情况下 $k = 5$,即可以将少数类扩充 100%,200%……500%。将新产生的合成样本加入到第二个新的训练集 T_2 中训练决策树算法 C4.5,由此得到二次分类器 C_2。

3) 对于待分类样本,首先利用初分类器 C_1 对其进行初分类,若被分类为多数类,则将其标记为多数类样本;若被分类为少数类,则利用二次分类器 C_2 对其进行再次分类。

综上所述,在 OSPC 算法中,我们训练了两个分类器:初分类器 C_1 和二次分类器 C_2。C_1 是以 T_1 为训练集的最近邻分类器,由于 T_1 是由多数类中对分类安全的“分类种子”集合 E 和少数类 S 组成的,因此,对待分类样本进行初分类时,若被分类为多数类,根据最近邻思想,它靠近多数类的“分类种子”,我们认为分类结果可信,则将其标记为多数类;若被分类为少数类,则它可能是少数类样本或者是离分类边

界相对较近的多数类样本，我们将利用二次分类器 C_2 对其进行再次分类。C_2 是 T_2 经过过抽样之后训练得到的决策树分类器，在 T_2 中，由于多数类样本集合 $H \subset M$，因此，与原始训练集 T 相比，T_2 中样本之间的不均衡程度减轻了，而且，少数类样本集合 S 经过过抽样之后，训练出的分类器对提高少数类的分类性能更加有利。

上述做法的目的是：初分类时，在保证少数类样本不被误分类的情况下，尽可能多地保留多数类的信息；二次分类时有效地提高少数类的分类性能。为了实现上述目的，训练集 T 中多数类 M 中的“分类种子”具有关键性的作用。如果“分类种子”选的太多，对待分类样本进行初分类时，少数类样本的误分类率就会增大，从而影响少数类的分类性能；如果“分类种子”选的太少，被 T_1 误分类的多数类样本集合 H 就会增大，从而不能更好地减轻 T_2 中样本的不均衡程度，对待分类样本进行二次分类时，少数类的分类性能不会得到更大地提高。

3 实验比较与分析

本文实验使用如表 2 所示的数据集。其中，Vehicle 是两类不均衡数据集，其余的是多类数据集。我们分别将 Breast-w 的第“4”类、Segment 的第“5”类和 Glass 的第“2”类看作它们的少数类，将其他类合并为多数类。表 2 中的数据选自 UCI 数据库（加州大学欧文分校数据库）[10]，其中的数据集通常作为机器学习与模式识别的评价数据集。

表 2 数据集的描述

数据集	样本个数	属性个数	类标号（少数类：多数类）	少数类的比例(%)
Breast-w	699	9	4: remainder	34.5
Segment	2310	19	5: remainder	14.3
Glass	214	9	2: remainder	35.5
Vehicle	846	18	1:0	23.5

对合成少数类过抽样算法 SMOTE、欠抽样方法（under-sample）与本文提出的基于初分类的过抽样算法 OSPC 进行了比较。其中，在欠抽样方法中去掉的是多数类中离分类边界相对较远的“分类种子”集合 E；被 SMOTE 和欠抽样方法处理过的训练集用决策树算法 C4.5 来验证其效果；对数据采用十交叉验证。此外，在 OSPC 算法的训练过程中，“分类种子”的选取方法如下：初分类时，在保证少数类样本的误分类率不超过 5% 的情况下，尽可能多地正确分类多数类样本。

图 3 中，因为欠抽样方法中少数类样本没有经过任何处理，它的 F-value 值与横坐标无关。但是，为了便于比较，将其画成直线。此外，SMOTE 和 OSPC 的实验结果曲线上横坐标为 0 时对应的 F-value 值为少数类的样本集合 S 没有被过抽样的情况。

从图 3 可以看出，在少数类的分类性能方面，基于初分类的过抽样算法 OSPC 与欠抽样方法相比要好的多，而且与过抽样算法 SMOTE 相比也有很大的提高。在 Breast-w 中，OSPC 的少数类的 F-value 的平均值比 SMOTE 的平均值和 under-sample 的 F-value 分别提高了 2.3% 和 7%；在 Segment 中分别提高了 4.4% 和 11%；在 Glass 中分别提高了 6.2% 和 8.4%；在 Vehicle 中分别提高了 2.4% 和 59%。

表 3 是 SMOTE、under-sample 和 OSPC 对数据集分类时多数类的 F-value 值。可以看出，在多数类的分类性能方面，OSPC 比 SMOTE、under-sampling 与 C4.5 有一定程度的提高。这说明，基于初分类的过抽样方法 OSPC 利用第一个训练集 T_1将多数类的信息有效地保留了下来。

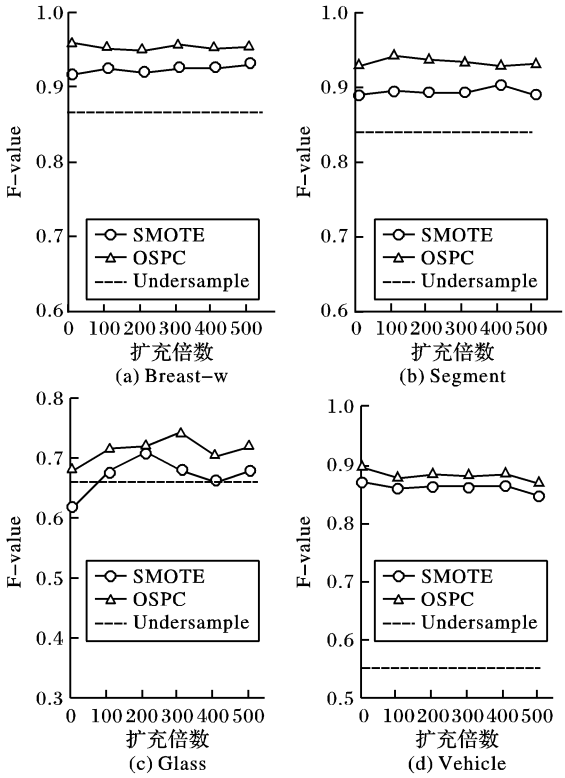

图 3 三种算法对数据集分类时少数类的 F-value 值

表 3 SMOTE、under-sample 和 OSPC 对数据集分类时多数类的 F-value 值

Methods		Dataset			
		Breast-w	Segment	Glass	Vehicle
C4.5		0.9551	0.9808	0.8056	0.9605
Under-sample		0.9148	0.9697	0.8281	0.7316
SMOTE	100%	0.9593	0.9815	0.8218	0.9556
	300%	0.9591	0.9801	0.8104	0.9563
	500%	0.9611	0.9799	0.8104	0.9508
OSPC	100%	0.9734	0.9896	0.8500	0.9611
	300%	0.9756	0.9884	0.8421	0.9626
	500%	0.9744	0.9879	0.8199	0.9577

综上所述，实验结果证实了基于初分类的过抽样算法 OSPC 的有效性。首先，对未知类标号的样本进行分类时，由表 3 可以看出，与欠抽样方法不同，OSPC 算法没有丢掉多数类的有用信息，而是由初分类器 C_1 在少数类样本不被错分的条件下，将其尽可能多地保留了下来；其次，从图 3 可以看出，与过抽样算法 SMOTE 相比，虽然少数类的过抽样倍数是一样的，但是 OSPC 的少数类的分类性能比 SMOTE 有了较大的提高。这是因为，训练集 $T\_2 = \{S, H\}$ 中，集合 H 是原始训练集中多数类的样本集合 M 的子集，所以 T_2 中样本的不均衡程度已经得到了减轻，并且在此基础上，继续对少数类进行过抽样，由此训练出来的二次分类器 C_2 肯定能更大程度地提高少数类的分类性能。

4 结语

不均衡数据集普遍存在于机器学习与模式识别的许多实际应用领域中，而且，少数类样本被错误分类所带来的损失要

比多数类样本被错误分类大的多。因此,如何有效地提高少数类的分类性能是这些领域追求的目标。本文提出的基于初分类的过抽样算法 OSPC,解决了已有方法中存在的不足,利用初分类与二次分类,既保留了多数类的有用信息,又更大程度地提高了少数类的分类性能。

我们将进一步研究如何更好地选取多数类的“分类种子”和自适应地确定“分类种子”的数量。此外,将特征选择方法与本文方法融合也是我们的研究方向。

参考文献:

[1] WEISS GM. Mining with rarity: A unifying framework[J]. Chicago, IL, USA, SIGKDD Explorations, 2004, 6(1): 7 - 19.

[2] CHAWLA NV, BOWYER KW, HALL LO, *et al.* SMOTE: Synthetic Minority Over-Sampling Technique[J]. Washington, USA, Journal of Artificial Intelligence Research, 2002, 16: 321 - 357.

[3] KUBAT M, MATWIN S. Addressing the Course of Imbalanced Training Sets: One-sided Selection[A]. Proceedings of the Fourteenth International Conference on Machine Learning[C]. San Francisco, CA, 1997. 179 - 186.

[4] BATISTA GEAPA, PRATI RC, MONARD MC. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data[J]. Chicago, IL, USA, SIGKDD Explorations, 2004, 6(1): 20 - 29.

[5] JOSHI M, KUMAR V, AGARWAL R. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements[A]. First IEEE International Conference on Data Mining[C]. San Jose, CA, 2001.

[6] WU G, CHANG EY. Class-Boundary Alignment for Imbalanced Dataset Learning[A]. Workshop on Learning from Imbalanced Datasets (ICML 03)[C]. Washington DC, 2003. 49 - 56.

[7] HUANG KZ, YANG HQ, KING I, *et al.* Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine[A]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition[C]. 2004.

[8] MANEVITZ LM, YOUSEF M. One-class SVMs for document classification[J]. Journal of Machine Learning Research, 2001, 2(2): 139 - 154.

[9] 边肇祺, 张学工. 模式识别[M]. 第 2 版. 北京: 清华大学出版社, 2000.

[10] BLAKE C, MERZ C. UCI Repository of Machine Learning Databases[DB/OL]. http://www.ics.uci.edu/~mlearn/~MLRepository.html, 2005.

关于征集中国计算机事业五十周年大事记的通知

为纪念中国计算机事业创建五十周年,中国计算机学会决定编辑出版“中国计算机事业五十周年大事记”。在“大事记”的基础上,由学会选评中国计算机事业发展历程中的五十件大事。编辑出版完成后,在今年 10 月下旬举行的“纪念中国计算机事业五十周年”活动上发布。

各相关单位或个人均可书面提供"大事记"的内容。“大事记”将反映对中国计算机事业发展有重要影响的事件、项目、发明或成果,"大事记"记录中还包括与该事件相关的主要单位和/或主要人士。

“大事记”提供者须完整填写下表,电子邮件发至:ccf@ict.ac.cn;原件同时寄至:北京 2704 信箱中国计算机学会,100080,注明“大事记”字样,也可传真至:010 - 6252 7485。

事件征集时间范围:1956 年 ~ 2005 年

征集截止日期:2006 年 6 月 30 日

中国计算机学会 2006 年 3 月 14 日

“大事记”推荐表

<table>
<tr><td>1 事件名称</td><td colspan="5"></td></tr>
<tr><td>2 起始日期</td><td colspan="2"></td><td>3 结束日期</td><td colspan="2"></td></tr>
<tr><td rowspan="4">4 参与该事件
主要单位</td><td colspan="5">单位 1:</td></tr>
<tr><td colspan="5">单位 2:</td></tr>
<tr><td colspan="5">单位 3:</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td>5 主要人士</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">6 提供者资料
(个人)</td><td colspan="3">姓名:</td><td colspan="2">现工作单位:</td></tr>
<tr><td colspan="3">任职:</td><td colspan="2">电话:</td></tr>
<tr><td colspan="3">Email:</td><td colspan="2">签字:</td></tr>
<tr><td rowspan="4">7 提供者资料
(单位)</td><td>单位名称</td><td colspan="2"></td><td colspan="2" rowspan="4">单位盖章</td></tr>
<tr><td>联 系 人</td><td colspan="2"></td></tr>
<tr><td>电　　话</td><td colspan="2"></td></tr>
<tr><td>E-mail</td><td colspan="2"></td></tr>
</table>

注:4:限 3 个单位;5:可选,限 5 位;6 和 7 选其一