

一种基于 HDDT 集成的多类不平衡学习方法

钱 祺, 姜 远

(南京大学 计算机软件新技术国家重点实验室, 江苏 南京 210093)

摘 要: 在很多真实世界问题中, 不同类别的数据样本往往有显著的不平衡性, 即大类的样本远多于小类. 对类别不平衡样本进行学习, 是目前国内外数据挖掘和机器学习领域的研究热点之一. 以往对不平衡样本学习的研究主要针对二分类问题进行, 由此针对多分类问题, 提出一种基于 HDDT 决策树集成的多类不平衡学习方法. 实验表明, 该方法可以有效地对多类不平衡问题进行学习.

关键词: 机器学习; 数据挖掘; 类别不平衡学习; 多分类

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2011)10-0201-03

A Multi-Class Imbalance Learning Method Based on HDDT Ensemble

QIAN Qi, JIANG Yuan

(National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, China)

Abstract: In many real world applications, the number of examples from different class is significantly different, which means the number of examples in major class is much larger than that of minor class. Therefore, learning from imbalanced data set has received much attention of machine learning and data mining community. Considering that most of previous research focus on binary class problem, this paper proposes a multi-class imbalance method based on HDDT ensemble. Empirical study shows that the method is effective for multi-class imbalance learning.

Key words: machine learning; data mining; class-imbalance learning; multi-class classification

1 引言

类别不平衡的现象在现实数据中普遍存在. 当某些类的样本数远大于另外一些类时, 传统的以最小化错误率为目标的机器学习算法就会忽略掉小类的影响而追求高的分类正确率^[1]. 但是, 实际上, 小类的分类精度可能远比大类的重要, 忽略小类可能造成重大的损失. 因此, 类别不平衡学习具有很重要的现实意义.

类别不平衡学习的基本解决方案是 Rescaling 方法^[2], 通过改变样本的分布来消除类别不平衡的影响. Rescaling 有多种实现^[2], 并由此产生了众多的学习方法, 常见的包括上采样(oversampling)^[1]与下采样(undersampling)^[1]. 上采样通过复制小类的样本, 增加小类的样本数达到平衡; 下采样通过对

大类样本采样, 减少大类的样本数以达到平衡; 样本加权的方法通过对小类样本赋以更高的权重来增加小类的影响, 从而平衡类别之间的关系; 阈值移动通过将决策平面向小类一方移动, 从而平衡不同类别对决策的影响. 基于这些 Rescaling 策略, 产生了很多有效的类别不平衡学习方法.

以往对类别不平衡学习方法的研究主要考虑二分类问题, 而在真实世界问题中, 多分类问题随处可见, 因此, 有必要对多类不平衡问题进行研究. 本文提出了一种基于 HDDT 决策树集成的多类不平衡学习方法. 实验表明该方法取得了较好的效果.

2 本文方法

HDDT 决策树^[3]是处理二分类问题的决策树

收稿日期: 2011-04-22; 修回日期: 2011-06-25

基金项目: 国家自然科学基金(60975043); 中央高校基本科研业务费专项资金(1115020204, 1116020206)

算法,它采用 Hellinger 距离作为分枝条件.假设和是两个不同的连续分布,是参数,那么,Hellinger 距离可以定义为^[4]

$$d_H(P, Q) = \sqrt{\int_0^1 (\sqrt{P} - \sqrt{Q})^2 d\lambda}$$

Cieslak 等人将其改写成适用于决策树分枝条件的形式^[3],即:

$$d_{H(x_+, x_-)} = \sqrt{\sum_{j=1}^p \left(\sqrt{\frac{x_{+j}}{x_+}} - \sqrt{\frac{x_{-j}}{x_-}} \right)^2}$$

式中, x_+ 代表正类样本, x_- 代表负类样本, x_{+j} 代表对于当前属性,具有第种属性值的正类样本, p 表示当前属性的属性值的总个数.可以发现,对于 Hellinger 距离,类别的数量互相之间没有影响,因此,更适合类别不平衡学习.

需要说明的是,由于 HDDT 是为离散属性设计的分类方法,因此,对于连续属性,HDDT 算法采用和 C4.5 类似的处理过程:首先将属性值进行排序,然后将每个属性值作为阈值,将数据集分成两部分,计算 Hellinger 距离.最后,选择性能最好的阈值作为该属性值的阈值.

对于有 c 类的多分类问题,每个样本 x 都 $y \in \{1, \dots, c\}$ 有与之对应的类标记.首先,通过采用一对一的策略,即选取任意两个类配对,将这个 c 类的多分类问题降解($\frac{c}{2}$) 为个简单的二分类问题.在这些二分类问题上,分别训练 HDDT 决策树作为基分类器.最后,得到一个有着 $c(c-1)/2$ 个基分类器的集成.预测时,通过每个基分类器对样本的分类结果进行投票,得到最高票数的类标记即为样本的预测结果.

图 1 中总结了本文算法,图 2 中总结了通过得到的决策树集成进行预测的过程.

3 实验

为了直观地展示本文方法的有效性,我们在一类有三类的两维人工数据集上进行对比实验.分类器是两层的决策树图 3 总结了实验结果.

```

Algorithm: HDDT Ensemble
Input: Training set T
Index=1
for each pair of class i and j do
  Binary(Index)
  = HDDT( $\{x_p | y_p = i\} \cup \{x_q | y_q = j\}$ )
  Index++;
end for
Return: Binary
  
```

图 1 本文方法

```

Algorithm: Predict by voting
Input: Test data x, HDDT ensemble Binary
for each class i do
  vote(i)=0
end for
Index=1;
for each pair of class i and j do
  label=x predicted by Binary(Index)
  vote(label)++;
  Index++;
end for
Return: arg max{vote(i)}
  
```

图 2 预测过程

图 3(a)是原始的多分类数据集,可以发现,该数据集中大类的样本数远远大于小类的样本数.图 3(b)是采用两层的 C4.5 算法在该数据集上的结果.图 3(c)~(e)是采用 C4.5 作为基分类器的集成算法在该数据集上的结果.图中展示了分解后的二分类问题上的两层 C4.5 算法的结果.可以发现,分解之后的二分类问题,用相同的层数的决策树,可以比原问题更好的得到解决.但是图 3(d)也说明,在 C4.5 算法中,由于以最小化错误率为目标,小类的影响很容易被忽略.图 3(f)~(h)是以 HDDT 作为基分类器的集成算法在该数据集上的结果.以图 3(g)为例, HDDT 在二分类问题中,侦测到了小类,并且取得了很好的分类效果.这说明, HDDT 比 C4.5 更适合作为集成算法的基分类器.

实际实验采用了 10 个来自 UCI 的数据集^[5],表 1 总结了实验所使用的数据集信息.实验采用 5 次 2 倍交叉验证(fold cross-validation).传统的分类精度不适合评价类别不平衡算法的性能,因此, AUC(Area Under Roc Curve)常被用作评价准则^[6].由于本文考虑的是多分类问题,因此使用 MAUC^[7]作为评价准则.采用 HDDT 和 C4.5 作为基分类器的集成算法分别标记为“HDDT-E”和“C45-E”,对比方法有上采样方法(SMOTE)^[4],下采样方法(随机去除大类样本),样本加权方法(instance weight)以及原始的 C4.5 算法,并分别记为“SMOTE”,“Under”,“IW”和“C4.5”.对于上采样和下采样方法,分别通过增加小类样本和减少大类样本,使得所有类具有相同的样本数.对于样本加权方法,通过设置样本的权重为该类样本数的倒数,使得小类的样本具有更大的代价.实验中所有对比方法采用不剪枝的 C4.5 决策树作为分类器.

实验结果表明 SMOTE, Under 和 IW 都是在二分类问题上有效的类别不平衡算法,但是可以发

现,在多分类的情况下,这些算法的表现都不够理想. 在所有数据集上,都不如 HDDT-E 和 C45-E,在一些数据集上,甚至不如最简单的 C4. 5 算法. 这说明,直接利用针对二分类问题设计的方法来解决多类不平衡问题难以得到好的效果.

4 结束语

本文针对多类不平衡问题,提出了一种基于 HDDT 决策树集成的学习方法,实验表明,该方法可以有效地对多类不平衡问题进行学习.

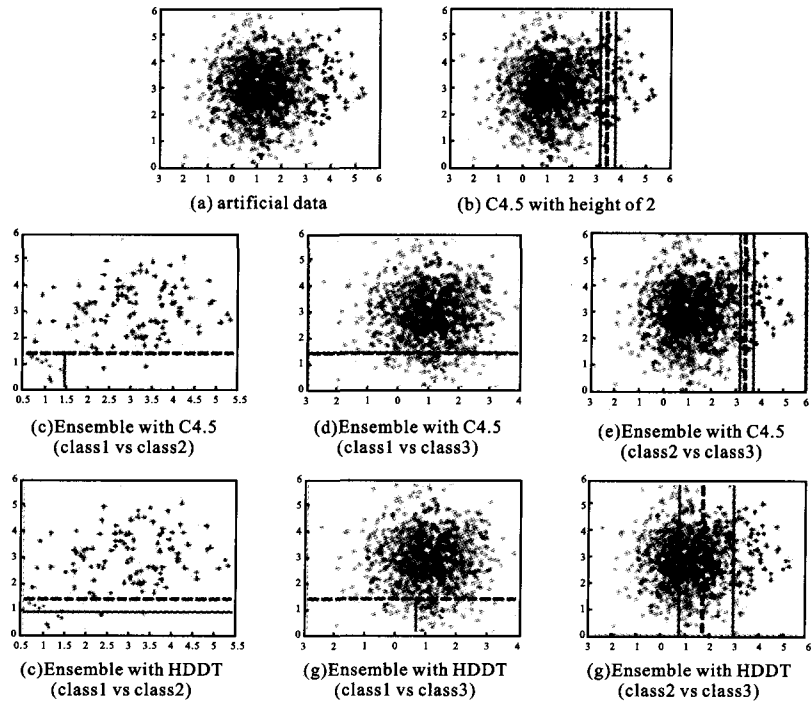


图 3 C4. 5 算法与分别以 C4. 5 和 HDDT 为基分类器的集成算法在人工数据集上的比较

表 1 实验数据集

数据集	样本个数	属性个数	类别个数	类别分布
abalone	4 177	8	4	[448; 2 282; 487; 960]
anneal	898	31	5	[8; 99; 684; 67; 40]
artificial	5 109	7	5	[2 006; 999; 500; 300; 1 304]
car	1 728	6	4	[1 210; 384; 69; 65]
cmc	1 473	9	3	[629; 333; 511]
hypothyroid	3 772	27	3	[3 481; 194; 97]
satellite	6 435	36	6	[1 533; 703; 1 358; 626; 707; 1 508]
segment	2 310	19	3	[330; 660; 1 320]
vehicle	846	18	3	[199; 218; 429]
yeast	1 484	8	9	[244; 463; 429; 163; 51; 44; 35; 30; 25]

-257.

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] Zhou Z H, Liu X Y. On multi-class cost-sensitive learning. Computational Intelligence, 2010 26(3), 232

[3] Cieslak D A, Chawla N V. Learning decision trees for unbalanced data [C]// proceedings of the 2008 European Conference on Machine Learning and knowledge Discovery in Databases - part I. Berlin, Heidelberg: springer-Verlay, 2008:241-256.

出采用分而治之策略的算法 DRMET,使之适用于稀疏型张量挖掘;该算法不仅继承了 MET 内存开销小的优点,同时在计算过程中巧妙地避免了因规模较大所造成的维数灾难问题,提升了算法执行效率,克服了 MET 挖掘运行时间过长缺陷;最后我们利用数据集测试了两款算法(MET 和 DRMET)的运行效率并加以比较分析,实验证明:所设计的新算法是高效、充分可适用的,将 MET 执行效率提升了约 1.86 至 15.85 倍。接下来,我们将要做的工作是考虑如何进一步提升其性能,尤其是当所挖掘张量并非方形张量时该如何使之适应。

参考文献:

- [1] Sun J, Papadimitriou S, Yu P. Window-based tensor analysis on high-dimensional and multi-aspect streams. [C]// proceedings of the 6th International Conference on Data Mining. Washington, DC, USA; IEEE Computer Society, 2006: 1076-1080.
- [2] Henrion R. Body diagonalization of core matrices in three-way principal components analysis: Theoretical bounds and simulation[J]. Journal of Chemometrics, 1993 (7): 477-494.
- [3] Carroll J D, Chang J J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition[J]. Psychometrika, 1970, 35(2): 283-319.
- [4] Tucker L R. Some mathematical notes on three-mode factor analysis[J]. Psychometrika, 1966, 31(2): 279-311.
- [5] Kolda T G, Sun J. Scalable tensor decompositions for multi-aspect data mining[C]// proceedings In ternational Conference on Data Mining. Washington, DC; IEEE, 2008.
- [6] Kolda T G, Bader B W. Tensor decompositions and applications[J]. SIAM Review, 2008(10): 1046-1092.
- [7] Wang H C, Wu Q, Shi L, et al. Out-of-core Tensor Approximation of multi-dimensional matrices of visual data[J]. SIGGRAPH, 2005, 24(3): 340-348.
- [8] Hitchcock F L. The expression of a tensor or a polyadic as a sum of products[J]. Journal of Mathematics and Physics, 1927(6): 164-189.
- [9] Bader B W, Kolda T G. Efficient MATLAB computations with sparse and factored tensors[J]. SIAM Journal on Scientific Computing, 2007, 30(1): 205-231.

作者简介:

周溜溜 男, (1987-), 硕士. 研究方向为数据挖掘机器学习.

(上接第 203 页)

- [4] Kailath T. The divergence and bhattacharyya distance measures in signal selection[J]. IEEE Transactions on Communication Technology, 1967, 15(1): 52-60.
- [5] Asuncion A, Newman D. UCI machine learning repository [EB/OL]. [2011-04-10]. <http://www.ics.uci.edu/mllearn/MLRepository.html>.
- [6] Bradley A P. The use of the area under the roc curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(7): 1145 - 1159.
- [7] Hand D J, Till R J. A simple generalisation of the area under the roc curve for multiple class classification problems [J]. Machine Learning, 2001, 45(2): 171 - 186.

作者简介:

钱 祺 男, (1986-), 硕士. 研究方向为机器学习与数据挖掘;

姜 远 女, (1976-), 博士, 副教授. 研究方向为机器学习与数据挖掘.