

基于非均衡数据集的代价敏感学习算法比较研究

谷 琼, 袁 磊, 熊启军, 宁 彬, 李文新

(襄樊学院 数学与计算机科学学院, 湖北 襄阳 441053)

摘 要: 大多数非均衡数据集的研究集中于重构数据集或者代价敏感学习, 针对数据集类分布非均衡和不相等误分类代价往往同时发生这一事实, 在简要回顾代价敏感学习理论和现有学习算法的基础上, 将所提出的自适应混合重取样算法, 与基于最小误分类代价的 MetaCost 算法分别进行实验比较, 实验表明所提出算法在代价敏感学习中具有一定的优势, 实验结果显示非均衡类对代价敏感学习算法性能产生较大影响, 当样本类别差异较大时, 用样本类空间重构的方法可以得到较好的分类效果。

关键词: 分类; 非均衡数据集; 混合重取样; 代价敏感学习

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2011)08-0146-04

A Comparative Study of Cost-Sensitive Learning Algorithm Based on Imbalanced Data Sets

GU Qiong, YUAN Lei, XIONG Qi-jun, NING Bin, LI Wen-xin

(Faculty of Mathematics & Computer Science, Xiangfan University, Xiangyang 441053, China)

Abstract: Most studies on the imbalanced data set classification focused on discussion of re-sampling or cost-sensitive learning systems themselves, however, the fact that imbalanced class distribution and misclassification errors cost unequally always occurring simultaneously was neglected. On the basis of analyzing the theory and algorithm of cost-sensitive learning, a novel hybrid re-sampling technique based on Automated Adaptive Selection of the Number of Nearest Neighbors in order to solve the misclassification problem of imbalanced data set is proposed. We compared hybrid re-sampling algorithm and MetaCost algorithm. Experiment results show that the proposed method can improve the classification accuracy and decrease the misclassification cost effectively. The experimental results confirm that this algorithm is superior to the traditional algorithms as for dealing with the imbalanced problem.

Key words: classification; imbalanced dataset; Hybrid Re-sampling; Cost Sensitive Learning

1 引言

目前, 机器学习已广泛用于工商企业以及科学研究等各个领域, 在这些应用中, 标准机器学习算法通常关注准确率, 只要提高算法预测的准确率就意味着提高了分类性能。但随着对实际应用的深入研究, 非均衡类数据分类问题大量出现, 即数据集中一类样本的数量远远超过其他类, 其中占大多数的样本被称为多数类, 而数量很小一部分的称为少数

类。同时, 人们发现有些问题不同类别样本的错分代价相差很大, 当用传统分类器虽然可得到较高的准确率, 但大量少数类样本却无法被准确识别。代价敏感学习(Cost Sensitive Learning, CSL)考虑非均衡类数据中正确识别少数类比多数类具有更大价值的事实, 根据不同类型的错分代价, 基于总体错分代价最小化来设计分类器, 在一定程度上解决了上述问题。

收稿日期: 2011-05-12; **修回日期:** 2011-06-28

基金项目: 国家自然科学基金项目(61075063); 国家高技术研究发展计划(“八六三”计划)(2009AA12Z117); 湖北省自然科学基金项目(2010CDB05201); 湖北省教育厅中青年项目(Q20112604)

2 代价敏感学习算法及评价指标

代价敏感学习算法针对不同类型的错分代价引入不同的惩罚因子,并基于总体误分代价最小化的原理来设计分类器.以二类分类情况为例,误分类代价所用的代价矩阵如表 1 所示^[1].

表 1 代价矩阵

真实类别	预测类别	
	预测为正类样本	预测为负类样本
正类样本	$\text{Cost}(P, P) = 0$	$\text{Cost}(P, N)$
负类样本	$\text{Cost}(N, P)$	$\text{Cost}(N, N) = 0$

其中, $\text{Cost}(i, j)$ 表示将样本 i 预测为 j 类时的代价

常用的代价敏感评价指标有:分类时产生的总的错误代价 (Total Costs, TC)、平均误分类代价 (Average Costs, AC) 等. TC 和 AC 分别为

$$\text{TC} = \text{FP} \times \text{Cost}(N, P) + \text{FN} \times \text{Cost}(P, N) \quad (1)$$

$$\text{AC} = \frac{\text{FP} \times \text{Cost}(N, P) + \text{FN} \times \text{Cost}(P, N)}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

代价敏感学习算法的研究集中在以下两方面^[2]:

(1) 重构训练集的方法,根据样本的不同错分代价给训练集中的每一个样本加权,接着按权重对原始样本集进行重构. Friedman^[3] 提出:找出类别分布使其倾向于代价高的类别并改变训练集.改变原始训练样本集的分布重新构建训练样本集,再用基于最小错误率的分类器作为基分类器进行分类实现. Elkan^[4] 曾提出关于改变正类和反类占总样本数的比例来构造较优的代价敏感学习算法.

(2) 通过改进分类器模型的内部结构,使其基于最小代价的代价敏感分类器.在传统分类算法中引入代价因子,对小样本赋予较高的代价,大样本赋予较小的代价,以此来平衡样本之间的数目差异^[5]. Fan^[6] 等人提出的 AdaCost 算法在 Boosting 算法迭代过程中检查错分代价较大的样本是否被错分,若被错分则其权值被较大增加,否则被减少然后进行下一轮分类. Maloof^[7] 指出非均衡数据集分类和不等误分代价学习可以使用同样的方法,但他并没有深入调查二者之间的影响.本文围绕非均衡数据集两类 CSL 解决方案展开比较研究.

3 混合重取样算法

本文设计一种基于重构数据集的解决方案自适应选择近邻的混合重取样方法 (ADSNNHRS),其

算法流程结构如图 1 所示.算法中对数据样本的处理分为两个阶段,先根据样本集内部的真实分布特性,自己选择调整少数类的近邻,在近邻候选集合中去除距离较远的少数类样本,并允许部分距离少数类样本较近的多类样本参与到新样本的生成过程中,对于近邻的不同情况采取不同的生成方法,从而达到控制少数类样本近邻区域、提高合成样本质量的目的;欠取样部分对合成之后的数据集去掉多数类中的冗余实例或边界上的噪音数据,在此基础上再进行分类和预测.该算法一方面通过自适应选择近邻的方法增加少数类样本的方式强调了少数类,另一方面对多数类进行适当程度的欠取样,减少其规模,最终达到多数类和少数类样本在一定程度上的相对均衡,从而可以有效地处理非均衡数据分类问题,提高分类器的性能.

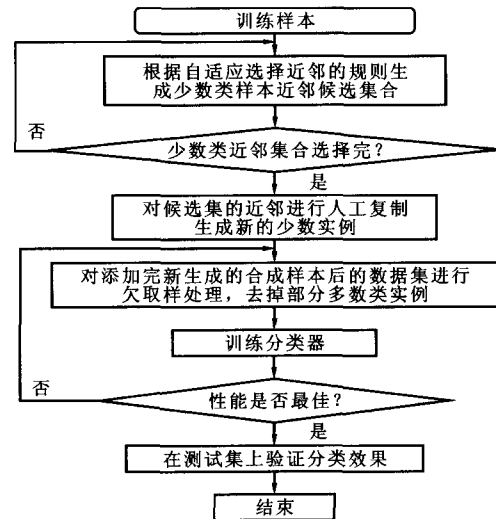


图 1 混合重取样算法处理流程图

4 两类算法的实验比较

现实生活中通常难于确定代价因子的数目.如在医疗诊断中,把癌细胞错分为健康细胞的代价很难确定具体设为多少才最合适.那什么情况下用样本类空间重构方法,什么情况下用调整算法的代价敏感学习算法,这两者对于同样的数据集使用有什么区别呢?下面将分别进行实验比较.

对于样本类空间重构方面,使用所提出的自适应选择近邻的混合重取样 ADSNNHRS 方法,先对数据集进行多次重取样,当数据基本均衡,然后再基于最小错误率的目标来进行分类,实验的基准算法选用 J48 决策树分类算法,并对数据集采用十折交叉验证法进行分类算法的评估.对于改进分类器的方法,选择 MetaCost 算法来对数据进行,仍选用 J48 为基准

学习算法,由于事先并不知道数据集的误分类代价因子,对于多数类,始终设置误分类代价因子为 1,对于少数类,根据数据集正负类数量相差的比例设置不同的代价因子,最大的代价因子设置成其与少数类数量的乘积是多数类数量的 2 倍,同样采用十折交叉验证法进行分类算法的评估. Sick、Ionosphere 数据集分别选自 UCI^[8] 的 Benchmark 数据集.

4.1 Sick 数据集实验

Sick 数据集共 3 772 个样本,包含 231 个 sick

实例,3 541 个 negative 实例,29 个预测属性和 1 个类别属性,其中 23 个名词型属性和 6 个数值型属性.

4.1.1 重取样的方法

Sick 数据集的 negative 类是 sick 类的 15 倍以上,采取 ADSNNHRS 方法将 Sick 数据集的 sick 类分别增加 2,4,8,16 倍,最终达到和 negative 类基本均衡的目的,得到的实验结果见表 2.

表 2 对 Sick 重取样后的分类结果

Sick	Class	TP	FP	Precision	F-Measure	G-mean	AUC
原始数据集	negative	0.997	0.242	0.984	0.991	0.869 3	0.929 4
	sick	0.758	0.003	0.946	0.841		
2 倍重取样	negative	0.993	0.082	0.989	0.991	0.954 8	0.983 8
	sick	0.918	0.007	0.946	0.932		
4 倍重取样	negative	0.992	0.017	0.995	0.993	0.987 5	0.996 7
	sick	0.983	0.008	0.969	0.976		
8 倍重取样	negative	0.991	0	1	0.995	0.995 5	0.997
	sick	1	0.009	0.984	0.992		
16 倍重取样	negative	0.99	0	1	0.995	0.995 0	0.997 6
	sick	1	0.01	0.99	0.995		

4.1.2 调整学习算法

测试中,TP 和 TN 的代价设为 0,negative 误分类代价因子(Cost Factor,CF)始终为 1,sick 类误分

类代价因子 CF 分别为 1,2,4,8,16,32. 选择 Meta-Cost 元代价算法来进行分类,得到的实验结果见表 3.

表 3 对 Sick 调整学习算法的分类结果

Sick	Class	TP	FP	Precision	F-Measure	G-mean	AUC
CF=1	negative	0.995	0.238	0.985	0.99	0.870 7	0.918 5
	sick	0.762	0.005	0.907	0.762		
CF=2	negative	0.99	0.199	0.987	0.989	0.890 5	0.922
	sick	0.801	0.01	0.841	0.82		
CF=4	negative	0.984	0.134	0.991	0.987	0.923 1	0.931
	sick	0.866	0.016	0.775	0.818		
CF=8	negative	0.975	0.121	0.992	0.984	0.925 8	0.933
	sick	0.879	0.025	0.7	0.779		
CF=16	negative	0.97	0.108	0.993	0.981	0.930 2	0.939
	sick	0.892	0.03	0.662	0.76		
CF=32	negative	0.97	0.108	0.993	0.981	0.930 2	0.939
	sick	0.892	0.03	0.662	0.76		

4.2 Ionosphere 数据集实验

Ionosphere 数据集共 351 个样本,包含 126 个 bad 实例,225 个 good 实例.34 个预测属性和 1 个类别属性,所有属性都是数值型属性.

4.2.1 重取样的方法

Ionosphere 数据集 good 类分别是 bad 类的 2 倍以上,采取 ADSNNHRS 方法将 Ionosphere 数据

集 bad 类分别增加 2,4 倍,再进行分类,最后得到的实验结果见表 4.

4.2.2 调整学习算法

假设 Ionosphere 数据集 good 的 CF 始终为 1,由于 Ionosphere 的不均衡程度并不高,这里我们将 bad 类误分类代价因子的设置高于两个类别之间的比例,CF 分别设为 1,2,4,8,得到的实验结果见表 5.

表 4 对 Ionosphere 重取样后的分类结果

Ionosphere	Class	TP	FP	Precision	F-Measure	G-mean	AUC
原始数据集	Good	0.96	0.302	0.85	0.902	0.818 6	0.878 5
	bad	0.698	0.04	0.907	0.789		
2 倍重取样	Good	0.943	0.075	0.905	0.924	0.934 0	0.956 1
	bad	0.925	0.057	0.955	0.94		
4 倍重取样	Good	0.868	0.022	0.944	0.904	0.921 4	0.953 1
	bad	0.978	0.132	0.946	0.962		

表 5 对 Ionosphere 调整学习算法的分类结果

Ionosphere	Class	TP	FP	Precision	F-Measure	G-mean	AUC
CF=1	Good	0.969	0.294	0.855	0.908	0.827 1	0.878 1
	bad	0.706	0.031	0.927	0.802		
CF=2	Good	0.942	0.27	0.862	0.9	0.829 3	0.895 8
	bad	0.73	0.058	0.876	0.797		
CF=4	Good	0.911	0.254	0.865	0.887	0.824 4	0.860 5
	bad	0.746	0.089	0.825	0.783		
CF=8	Good	0.787	0.151	0.903	0.841	0.817 4	0.854 9
	bad	0.849	0.213	0.69	0.762		

4.3 实验结果分析

从上面 2 个数据集的两种不同解决方案,可以看出,两种方法的分类效果都有不同程度地改进.对于 Sick 数据集,由于其不均衡程度非常高,两类数据之比接近 16 倍,其采用样本类空间重构的方法,分类效果提高十分明显,当数据集接近均衡时,少数类分类的误差为 0,同样对于 Sick 数据集,我们将少数类的错误分类代价设得比多数类的大很多时,分类效果也有明显的提高,当我们把代价因子设为 32 时,却得到了跟 16 同样的分类结果;而对于不均衡程度较低的 Ionosphere 数据集,当代价因子设为 8 时,产生的分类效果同 2 倍重取样的效果基本相当,但总体的分类效果比样本空间重构的效果差.这说明了非均衡类对代价敏感学习算法性能产生较大影响,并且在学习过程中,并不是代价因子越大分类性能越好,而是数据集越均衡,分类性能越好,在这种情况下,用样本类空间重构的方法可以得到较好的分类效果.只有当少数类

的错误分类代价比多数类大很多时,并且在处理不是很严重的非均衡类时,我们可以直接使用基于最小误分类代价的代价敏感学习方法.

5 结束语

代价敏感学习是机器学习的研究热点,和传统分类方法不同之处在于,更注重在分类时尽量避免产生高代价的错误,以达到分类时产生的错误总代价最低的目标.这种思想更加符合现实中的分类问题,并在入侵检测、银行信用、医疗诊断等众多领域呈现广泛的应用前景.非均衡类问题与误分类代价不同的学习完全相同,可以用相类似的解决途径来进行处理.目前处理非均衡类数据分类问题的方法虽然很多,但没有一种方法是最优的,迄今为止,没有一种算法可以在任何代价敏感分类问题上都取得最好结果,需要更多学者进一步去研究.

(下转第 153 页)

5 结束语

在传统机器人蒙特卡罗定位算法的基础上,针对粒子滤波中的粒子退化问题,利用扩展卡尔曼滤波来精确设计粒子滤波器,改进了粒子分布预测函数,减少了所需粒子数;采用MCMC方法进行重采样,并通过使用变概率分布函数边界的办法来增强粒子分散能力.仿真实验表明:提出算法仅采用较少粒子即可达到较高的位姿跟踪精度,同时有效粒子数也大为增加.特别当系统噪声较大而观测信息相对精确时,IRR-MCL更体现出了较传统MCL的精确性与鲁棒性.

参考文献:

- [1] Fox D, Burgard W, Dellaert F, et al. Monte carlo localization: efficient position estimation for mobile robots[C]// In: proceedings of National conference on Artificial Intelligence. Orlando, Florida, USA: AAAI/MIT Press, 1999: 343—349.
- [2] Thrun S, Fox D, Burgard W, et al. Robust monte carlo localization for mobile robots[J]. Artificial Intelligence, 2000, 128(1): 99—141.
- [3] Fox D. Adapting the sample size in particle filters through KLD—sampling[J]. International Journal of Robotic Research, 2003, 22(1): 985—1003.
- [4] Lenser S, Veloso M. sensor resetting localization for poorly modeled mobile robots[C]// Proceedings of IEEE International Conference on Robotics and automation, San Francisco, USA: IEEE, 2000: 1225—1103.
- [5] Ueda R, Arai T, Sakamoto K, et al. Expansion resetting

for recovery from fatal error in monte carlo localization—comparison with sensor resetting methods[C]// Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan: IEEE, 2004: 2481—2486.

- [6] Doucet A, Freitas N de, Gordon N. Sequential monte carlo methods in practice[M]. New York: Springer—Verlag, 2001.
- [7] Hu ShiQiang, Jing ZhongLiang. Overview of particle filter algorithm[J]. Control and Decision, 2005, 20(4): 361—365.
- [8] Deng XiaoLong, Xie JianYing, Wang LinSu. Improved particle filter based on current statistical model[J]. Control and Decision, 2005, 20(5): 567—570.
- [9] Wu ErYong, Xiang ZhiYu, Liu JiLin. Robust robot monte carlo localization[J]. Acta Automatica Sinica, 2008, 34(8): 907—911.
- [10] 宋宇, 孙富春, 李庆玲. 移动机器人的改进无迹粒子滤波蒙特卡罗定位算法[J]. 自动化学报, 2010, 36(6): 851—857.
- [11] 张恒, 樊晓平, 瞿志华. 基于多假设跟踪的. 移动机器人自适应蒙特卡罗定位算法[J]. 自动化学报, 2007, 33(9): 941—946.
- [12] van der Merwe R, de Freitas N, Doucet A, et al. The unscented particle filter, Technical Report CUED/F—INFENG/TR 380[R]. UK: Department of Engineering, Cambridge University, 2000.

作者简介:

杨居义 男, (1959—), 学士, 高级工程师, 副教授. 研究方向为智能控制与机器人、数字图像处理.

(上接第149页)

参考文献:

- [1] 凌晓峰, Sheng Victor S. 代价敏感分类器的比较研究[J]. 计算机学报, 2007(8): 1203—1211.
- [2] 赵会, 黄景涛, 谈书才. 最小二乘支持向量机的一种非均衡数据分类算法[J]. 微电子学与计算机, 2010, 27(4): 33—37.
- [3] Friedman J H, Olshen R A, Stone C J, et al. Classification and regression trees[M]. American Statistical Association: The Film House, 1986.
- [4] Elkan C. The foundations of cost—sensitive learning[C]// Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01). Washington DC, 2001: 973—978.
- [5] Ciraco M, Rogalewski M, Weiss G. Improving classifier utility by altering the misclassification cost ratio[C]//

the 1st International Workshop on Utility—based Data Mining. New York, 2005: 46—52.

- [6] Fan W, Stofol S, Zhang J X. Ada cost: misclassification cost—sensitive boosting[C]// Proc of the 16th Int'l Conf on Machine Learning. Slovenia: Bled, 1999: 97—105.
- [7] Maloof M. Learning when data sets are imbalanced and when costs are unequal and unknown[C]// Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets. Washington, DC, 2003.
- [8] The Center for Machine Learning and Intelligent Systems. UC Irvine machine learning repository[DB/OL]. [1989—01—01]. <http://archive.ics.uci.edu/ml/datasets.html>.

作者简介:

谷琼 女, 博士, 讲师, CCF 会员. 研究方向为数据挖掘、演化计算等.