

# 一种基于欠采样的不平衡数据分类算法

程险峰<sup>1</sup>, 李 军<sup>2,3</sup>, 李雄飞<sup>3</sup>

(1. 长春市公安局交通警察支队, 长春 130011; 2. 长春理工大学数学系, 长春 130022;  
3. 吉林大学符号计算与知识工程教育部重点实验室, 长春 130012)

**摘 要:** 针对不平衡数据学习问题, 提出一种基于欠采样的分类算法。对多数类样例进行欠采样, 保留位于分类边界附近的多数类样例。以 AUC 为优化目标, 选择最恰当的邻域半径使数据达到平衡, 利用欠采样后的样例训练贝叶斯分类器, 并采用 AUC 评价分类器性能。仿真数据及 UCI 数据集上的实验结果表明, 该算法有效。

**关键词:** 机器学习; 分类算法; 不平衡数据; 欠采样; 邻域

## Imbalanced Data Classification Algorithm Based on Undersampling

CHENG Xian-feng<sup>1</sup>, LI Jun<sup>2,3</sup>, LI Xiong-fei<sup>3</sup>

(1. Traffic Police Detachment, Changchun Public Security Bureau, Changchun 130011, China;

2. Dept. of Mathematics, Changchun University of Science and Technology, Changchun 130022, China;

3. Key Laboratory of Symbolic Computation and Knowledge Engineering for Ministry of Education, Jilin University, Changchun 130012, China)

**【Abstract】** Imbalanced Data Learning(IDL) problem is one of the research issues in machine learning. This paper presents a classification algorithm based on undersampling, which algorithm undersamples the majority examples, and retains the majority examples near the classify border. With the AUC as the optimization objectives. It chooses the most appropriate domain radius to balance the data set, and trains the Bayesian classifier by the use of the examples after undersampling. Using AUC as a measure of classifier performance evaluation, the experiments on simulation data and UCI data sets show that undersampling is effective.

**【Key words】** machine learning; classification algorithm; imbalanced data; undersampling; neighborhood

DOI: 10.3969/j.issn.1000-3428.2011.13.047

### 1 概述

在许多实际应用领域中存在数据不平衡情形, 例如信用欺诈、入侵检测、医疗诊断等。近年来, 不平衡数据学习(Imbalanced Data Learning, IDL)问题得到了机器学习研究者的广泛关注。在数据不平衡情况下, 多数类的信息占主导地位, 标准分类器往往忽视少数类的信息, 倾向于学习预测多数类, 造成分类器分类性能偏低。目前针对不平衡学习问题的主要解决方案包括基于采样的方法、基于 boosting 的算法、代价敏感学习、核学习和单一类别学习等<sup>[1]</sup>。本文提出一种新的数据平衡方法——利用欠采样删除远离边界的多数类样例, 但保留位于分类边界附近的多数类样例, 在相对平衡的训练数据集上构建贝叶斯分类器。

### 2 相关工作

#### 2.1 基于采样的不平衡数据学习

基于采样的方法是算法独立的, 不同的采样技术可以和 boosting 算法、核学习等分类算法结合。采样的主要目的是通过添加少数类样例(过采样), 或者移除多数类样例(欠采样)的方式平衡类分布信息<sup>[2]</sup>。研究表明, 利用平衡后的数据集, 可以获得比原始不平衡数据更优的分类性能<sup>[3]</sup>。

过采样方法通过添加或复制少数类样例, 调整类分布的不平衡程度, 其缺点是可能会产生较多的边际或噪音样例; 欠采样删除部分多数类训练样例, 使多数类信息不再占主导地位, 而少数类信息得到相对充分的表达, 但其缺点是可能引起信息丢失, 尤其是对寻找分类边界有益的信息。文献[4]对多数类样本进行粒划分, 并在局部支持向量和少数类样本上学习, 使支持向量机(Support Vector Machine, SVM)获得了

令人满意的泛化能力。

#### 2.2 不平衡数据学习的性能评价

在两分类的情形下, 将少数类视为正类, 多数类视为负类。在评估不平衡数据分类器时, 分类准确度、误差率等传统的评估度量不再有效, 例如, 在正负类不平衡比率为 5:95 的情况下, 即使将所有样例都判定成负类, 其准确率也高达 95%, 但该分类器在正类上的准确度却为 0。

不平衡数据分类的常用评价标准包括基于混淆矩阵的若干度量。图形化性能度量包括接收者操作特征(Receiver Operating Characteristics, ROC)曲线、precision-recall 曲线、cost 曲线等。数值化性能度量包括正确率、精确度(precision)、召回率(recall)、F-measure、gmean 和 ROC 曲线下方面积(Area Under ROC curve, AUC)<sup>[1]</sup>。混淆矩阵表示样例分类后的 4 种情况, 如图 1 所示。

	预测正类	预测负类
实际正类	True Positives(TP)	False Negatives(FN)
实际负类	False Positives(FP)	True Negatives(TN)

图 1 混淆矩阵

**基金项目:** 国家科技支撑计划基金资助项目(2006BAK01A33); 公安部重点科研基金资助项目(B 类)(20032252001); 吉林省科技发展计划基金资助项目(20070321, 20090704)

**作者简介:** 程险峰(1955—), 男, 高级工程师, 主研方向: 智能交通, 数据挖掘; 李 军, 副教授、博士; 李雄飞, 教授、博士生导师

**收稿日期:** 2011-02-25 **E-mail:** lijun.yq@163.com

利用混淆矩阵可派生出以下度量:

$$(1) \text{真正率(True Positive rate): } TP\text{rate} = \frac{TP}{TP + FN};$$

$$(2) \text{假正率(False Positive rate): } FP\text{rate} = \frac{FP}{TN + FP};$$

$$(3) \text{真负率(True Negative rate): } TN\text{rate} = \frac{TN}{TN + FP};$$

$$(4) \text{假负率(False Negative rate): } FN\text{rate} = \frac{FN}{TP + FN}.$$

ROC 曲线作为分类器评估的可视化技术<sup>[5]</sup>, 得到了广泛应用。在 ROC 曲线中,  $y$  轴表示真正率, 而  $x$  轴表示假正率。ROC 曲线上的每个点对应一个分类器模型。

令  $y$  表示  $TP\text{rate}$ ,  $x$  表示  $FP\text{rate}$ , 则 ROC 曲线下方面积 AUC 为  $\int_0^1 y dx$ 。AUC 是 ROC 曲线的量化表示, 本文采用 AUC 作为分类器评价度量。

### 3 基于欠采样的不平衡数据学习算法

#### 3.1 数据欠采样方法

人类对事物进行分类时, 对于容易区分的样例, 可以首先进行类别标记, 将其移除样本空间。然后对不容易区分的事物采取特殊的识别方法。通常远离类别边界的样例总是易于区分的, 这些样例对于找到分类的边界基本没有帮助, 而类别边界附近的样例通常是难于区分的。基于以上思想, 本文提出一种欠采样方法, 移除远离分类边界的样例, 保留边界附近的多数类, 使得数据集分布得到一定程度的平衡, 然后在平衡数据集上构建分类器, 给出分类边界。

**定义 1** 如果某个多数类样例的邻域内没有少数类样例, 则称该样例是  $\delta$  可去的。

**定义 2** 如果某个多数类样例的邻域内有少数类样例, 则称该样例是  $\delta$  不可去的。

平衡训练集的操作步骤如下:

(1) 对于给定多数类样例, 判定该样例是否是  $\delta$  可去的。

(2) 删除所有  $\delta$  可去样例, 由少数类样例和  $\delta$  不可去的多数类样例构成欠采样数据集。

在欠采样数据集上调用贝叶斯学习算法得到分类器。由于保留了存在于边界附近的样例, 因此可有效避免信息缺失。

图 2 是原始仿真数据示意图, 其中, “o” 表示多数类, “+” 表示少数类。显然, 多数远离分类边界的类样例可以首先移除。

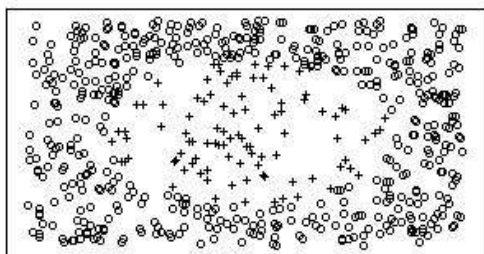


图 2 原始数据分布

在分类器训练时, 训练数据集由少数类样例和  $\delta$  不可去的多数类样例构成。

#### 3.2 样例间距离的计算方法及邻域半径 $\delta$ 的选择

##### 3.2.1 样例间距离的计算方法

由于样例的属性有连续型和离散型 2 种, 因此需要组合不同属性的距离度量。假设  $e_i$  和  $e_j$  为 2 个不同样例, 对不同类型属性, 采用不同方法计算样例在该类型属性下的距离(见图 1), 然后对不同类型距离进行加权, 得到样例间的距离<sup>[6]</sup>。

不同属性下的距离计算公式如下:

$$(1) \text{连续型: } d = \|e_i - e_j\|_2;$$

$$(2) \text{离散型(标称): } d = \begin{cases} 0 & \text{if } e_i(a) = e_j(a) \\ 1 & \text{if } e_i(a) \neq e_j(a) \end{cases};$$

$$(3) \text{离散型(序数): } d = |e_i - e_j| / n \quad (n \text{ 为值的个数}).$$

##### 3.2.2 邻域半径的选择

上述欠采样方法中对多数类样例的“移除”, 只是将其从样本空间中移除, 实际上是这些样例远离分类边界, 易于区分。因此, 并不是移除越多越好, 过小的邻域半径  $\delta$ , 将导致所有的多数类样例都被移除, 或者被保留的多数类样例太少, 在欠采样后的分类器训练集中, 这些原来的多数类样例反而占较小的比例, 这时无法找到恰当的分类边界。

图 3 所示为较小邻域半径对分类的影响。其中, 深色样例为  $\delta$  不可去样例; 浅色样例为  $\delta$  可去样例(图 4~图 5 同上)。由图 3 可见, 由于  $\delta$  不可去的多数类样例过少, 分类效果很差。

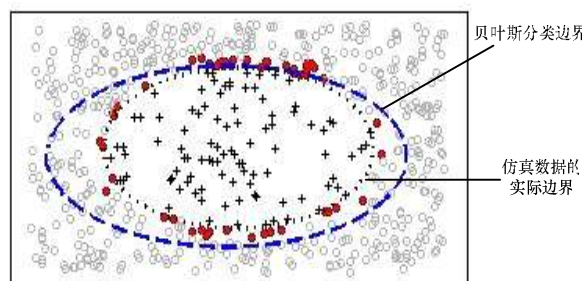


图 3 采用较小领域的分类边界

过大的邻域半径  $\delta$  会使  $\delta$  不可去多数类样例过多, 分类器训练数据集仍不够平衡, 从而影响分类效果, 图 4 所示为采用较大邻域半径对分类的影响。

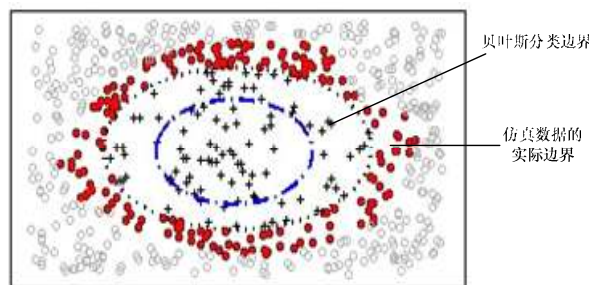


图 4 采用较大邻域的分类边界

综上所述, 邻域半径  $\delta$  的选择, 直接关系到算法的性能。本文算法采用 AUC 作为性能度量, 以优化 AUC 为目标, 求得最佳的邻域半径  $\delta$ 。

#### 3.3 算法步骤及分类结果

基于欠采样的不平衡数据分类算法如下:

**输入** 多数类别样例与少数类别样例, 备选的邻域半径  $\delta$  的集合  $\Delta$

**输出**  $h_{\delta}(X)$

(1) 扫描原始数据集  $O$ , 假设集合  $O_{\text{maj}}$  是多数类别样例集合, 标签为 -1, 令  $O_{\text{min}} = O - O_{\text{maj}}$  表示少数类别样例集合, 标签为 1。训练集  $T = O_{\text{min}}$ ,  $U(\text{example}, \delta)$  表示样例  $\text{example}$  的  $\delta$  邻域。

(2) for  $\delta \in \Delta$  do

for  $\text{example} \in O_{\text{maj}}$  do

if  $U(\text{example}, \delta)$  中有少数类样例 then 将  $\text{example}$  标记为  $\delta$  不可去的;  $T = T \cup \text{example}$ ;

endif

if  $U(\text{example}, \delta)$  中没有少数类样例 then 将  $\text{example}$  标记为  $\delta$  可去的;

```

endif
endfor
在训练数据集 T 上调用贝叶斯分类算法, 获得分类器  $h_{\delta}: X \rightarrow \{-1, 1\}$ ;
利用测试集, 计算分类器  $h_{\delta}$  的 AUC $_{\delta}$ 值;
endifor
(3) 计算具有最大 AUC 值的分类器所对应的邻域半径  $\delta^*$ ,  $\delta^* = \operatorname{argmax}_{\delta} \{AUC_{\delta}\}$ 

```

上述算法的分类结果如图 5 所示。可以看出, 基于欠采样方法的贝叶斯分类器分类边界更接近于真实分类边界。

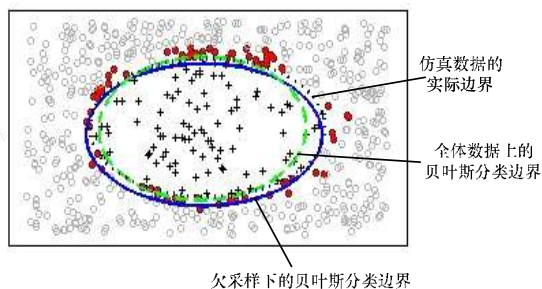


图 5 采用恰当邻域的分类边界

## 4 实验与分析

### 4.1 数据集及实验设置

为评估算法的性能, 实验选择 6 组具有不同实际应用背景的 UCI 数据。对含有多个类别的数据, 合并某些类别作为负类, 如表 1 所示。为保持类别分布一致, 利用分层抽样将原始数据分成训练集和测试集: 60%用于训练, 40%用于测试。实验基于 Matlab 2010a 软件环境, 选择朴素贝叶斯分类算法。

表 1 UCI 数据集

数据集	样例数目	类分布	属性个数	
			连续	离散
Segment	2 310	0.14:0.86	19	0
Breast-W	699	0.66:0.34	9	0
Abalone	731	0.06:0.94	7	1
Vehicle	846	0.23:0.77	18	0
Vowel	990	0.09:0.91	10	3
Satimage	6 435	0.09:0.91	33	0

### 4.2 实验结果及分析

实验结果如图 6 所示。可以看出, 基于欠采样的贝叶斯分类器性能优于在不平衡数据上贝叶斯分类器的性能。

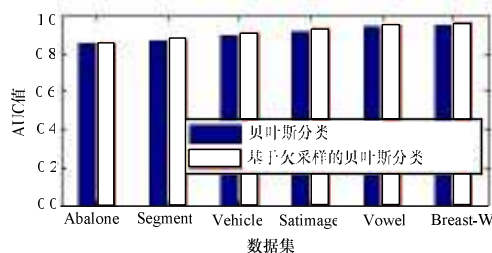


图 6 测试集上的 AUC 值

## 5 结束语

对于不平衡数据学习问题, 采样技术是常用方法之一, 合理的欠采样可以避免分类器对多数类的过度学习倾向。本文定义多数类  $\delta$  不可去样例和  $\delta$  可去样例的概念, 利用欠采样仅保留位于边界附近, 对分类有益的样例, 移除  $\delta$  可去样例。在得到欠采样后的训练数据集后, 利用贝叶斯分类算法训练分类器。以 AUC 作为性能评估度量, 通过优化 AUC, 得到最恰当邻域半径  $\delta$ 。仿真和实际数据集上的实验表明, 基于欠采样的分类器性能优于在原始数据集上的分类器。下一步研究将围绕基于数据的分形结构实现欠采样, 以便欠采样结果更真实反映类别边界附近的样例分布。

### 参考文献

- [1] He Haibo, Edwardo A. Learning from Imbalanced Data[J]. IEEE Trans. on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
- [2] Chawla N V, Japkowicz N, Kolcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets[J]. SIGKDD Explorations, 2004, 6(1): 1-6.
- [3] Batista G E A, Prati R C, Monard M C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.
- [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] Fawcett T. An Introduction to ROC Analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874.
- [6] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining[M]. Boston, Massachusetts, USA: Addison Wesley, 2005.

编辑 金胡考

(上接第 140 页)

## 6 结束语

本文利用图像关键特征点在攻击中不易丢失的特点, 提出一种新的算法。该算法利用 Harris 算法提取特征点, 以特征点为中心形成方形区域进行小波变换, 将水印信息嵌入到低频部分, 实验结果表明, 该算法具有较好的鲁棒性, 特别是剪切和压缩效果非常好。该方法可用于数字媒体的版权保护。仿真中得到以下结论: (1) 嵌入水印的方形区域不能太多, 即特征点不能选择太多, 否则会影响不可见性; (2) 特征点数量的多少取决于 Harris 算子窗口半径和设定的阈值; (3) 阈值的不同会影响特征点选择的速度, 从而影响整个算法的速度和计算量。但对于没有实时性要求的版权保护来说, 可以不予考虑。

### 参考文献

- [1] 李 健, 叶有培, 何春梅, 等. 基于 SIFT 特征点的抗几何攻击水印算法[J]. 计算机工程, 2009, 35(19): 170-171, 174.

- [2] Marco L. The Improbability of Harris Interest Points[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(6): 1141-1147.
- [3] Wang Xiangyang, Yang Yiping, Yang Hongying. Invariant Image Watermarking Using Multi-scale Harris Detector and Wavelet Moments[J]. Computers and Electrical Engineering, 2010, 36(1): 31-44.
- [4] Kutter M, Bhattacharjee S K, Ebrahimi T. Towards Second Generation Watermarking Schemes[C]//Proc. of the 6th Int'l Conf. on Image Processing. Kobe, Japan: [s. n.], 1999.
- [5] Harris C, Stephen M. A Combined Corner and Edge Detector[M]. Manchester, UK: [s. n.], 1988.
- [6] 李传目, 宋海明. 抗几何攻击的图像水印算法[J]. 郑州大学学报, 2009, 30(2): 75-79.

编辑 陈 文



