

标记样本的 Adaboost 算法

郑 诚, 张 瑞, 陈娟娟

(安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039;

安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘 要:提升(Boost)学习算法中,可以划分为多数提升和 Adaboost 两类。Adaboost 是目前比较流行的分类方法,目前在多媒体和计算机视觉领域得到了广泛的应用。文中介绍了 Adaboost 方法的原理与方法,通过在提升过程中对训练集中部分样本的标记,提出了一种新的 Adaboost 算法的训练方法,并且用实验数据对该方法进行验证。该方法通过对前一轮提升后权值较小的那部分样本作标记,增加了后一轮提升抽样的有效容量,从而使算法中的分类器能够更快速地关注那些很难分类的样本。

关键词: Adaboost 算法;提升;抽样;有效容量

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2008)07-0109-03

An Adaboost Algorithm with Sample Marked

ZHENG Cheng, ZHANG Rui, CHEN Juan-juan

(Edu. Dept. Key Laboratory of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China;

School of Computer Science and Technology of Anhui University, Hefei 230039, China)

Abstract: In multimedia and computer fields, Adaboost algorithm is the popular boost learning method for classification. It can be divided into two series; Boost by majority and Adaboost. The algorithm of Adaboost method was described in detail. A new training method has been raised through the marks on the part of training samples and proved by experiment. The former's effective capacity of sample has been improved by the new algorithm, which is made on the former samples with less weights. Thus, the classification algorithm can focus on those samples faster which are hard to classify.

Key words: Adaboost algorithm; boost; sample; effective capacity

0 引言

数据收集和数据存储技术的快速进步使得各组织机构可以通过积累得到海量数据。从而,提取有用的信息成为巨大的挑战。通常,由于数据量太大,无法使用传统的数据分析工具和技术处理它们。有时,即使数据集相对较小,由于数据本身的非传统特点,也不能使用传统的方法进行处理。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中的或人们事先不知道的,但又是潜在有用的信息和知识的过程。一般来说,数据挖掘是一个利用各种分析方法和分析工具在大规模海量数据中建立模型和发现数据间关系的过程,这些关系和模型可以用来

作出决策和预测,是在大型数据存储库中,自动地发现有信息的过程。数据挖掘将传统的数据分析方法与处理大量数据的复杂算法相结合。数据挖掘技术可以用来支持广泛的商务智能应用,如顾客分析、定向营销和欺诈检测等。通常,数据挖掘任务分为预测任务和描述任务两大类。其中,预测建模包含分类(classification)和回归(regression)两类子任务。两项任务目标都是训练一个模型,使目标变量预测值与实际值之间的误差达到最小。

分类作为数据挖掘的主要内容之一,主要是通过分析训练数据样本,产生关于类别的精确描述。分类就是依照所分析对象的属性分门别类、加以定义、建立组别。分类的关键是确定对数据按照什么标准或什么规则进行分类。因此,分类时首先根据属性特征,为每一类别找到一个合理的描述或模型,即确定分类规则,再根据规则对数据进行分类。分类的任务是确定对象属于哪个预定义的目标类,分类预测的是类别标号,即离散值。分类知识的应用范围广泛,包括信誉度鉴别、

收稿日期:2007-10-15

基金项目:国家自然科学基金资助项目(60475017);安徽省高等学校自然科学基金项目(2006kj055B)

作者简介:郑 诚(1964-),男,副教授,硕士生导师,主要从事数据挖掘、机器学习研究。

医疗诊断、性能测试和购物分析等。文中涉及的 Adaboost 算法即为分类方法中的一种^[1]。

1 Boost 方法概述

1.1 基本概念

样本:分类任务的输入数据是记录的集合,每条记录也称为实例或样例,用元组 (x, y) 表示,其中 x 是属性的集合,而 y 是样例的类标号。

分类:分类任务就是通过学习得到一个分类模型 f ,把每个属性集 x 映射到一个预先定义的类标号 y 。

训练集:由类标号已知的记录组成,使用训练集建立分类模型。

检验集:由类标号未知的记录组成,用于对分类模型的测试^[2]。

1.2 Boost 方法概述

Boost 由 Freund 和 Schapire 于 1990 年提出,是组合学习方法中最具有代表性的一种,提高了预测学习模型的预测精度。Boost 通过对训练集的操作以产生多个子预测模型,从而建立结合了投票原理的预测器集合。提升(Boost)是一个迭代的过程,用来自适应地改变训练样本的分布,使得简单分类器关注那些很难分的样本上。提升给每个训练样本赋一个权值,而且在每一轮提升过程结束时自动地调整权值。

Boost 方法主要是通过对样本集的操作来生成一系列的分类器,用来提高其他分类算法的精度^[3]。也就是将其他的分类算法放于 Boost 框架中,通过 Boost 框架对训练集的操作,得到不同的子训练集,每得到一个样本集就用该算法在该样本集上产生一个分类器,这样在给定训练次数 n 后,就可产生 n 个子分类器,然后 Boost 框架算法将这 n 个子分类器进行加权组合,产生一个最终的分类器,在这 n 个子分类器中,单个分类器的精度不一定很高,但它们组合后的结果有很高的精度,这样便提高了该类分类器的最终识别率。在产生单个分类器时,可用相同的分类算法,也可用不同的分类算法,这些算法一般为分类能力较弱的分类算法。

Adaboost 算法为 Boost 中的代表性算法。Adaboost 在训练集上维持一套概率分布,在每一轮的迭代中,Adaboost 在每个样本上调整这种分布,子分类器在训练集上的错误率被计算出来并以此在训练集上调整概率分布。权重改变的作用是在被错误分类的样本上增加更多的权重,在分类正确的例子上减少权重。通过单个分类器的加权投票建立最终分类器,每个子分类器按其训练集上的精度而加权。

文中在 Adaboost 的基础上提出了一种改进算法。该算法添加了对 Adaboost 算法中样本的标记,从而增

加了在算法提升过程中,通过抽样产生的训练集的有效容量。

2 Adaboost 算法

Adaboost 算法是一种分类算法,由 Yoav Freund 和 Robert E. Schapire 在 1995 年提出^[4]。算法的基本思想是:利用分类能力一般的简单分类器(weaker classifier),通过一定的方法进行提升(boost),最后生成一个分类能力很强的强分类器(strong classifier)。理论证明,只要单个简单分类器的分类能力比随机猜测要好,当简单分类器个数趋向于无穷时,强分类器的错误率将趋于零。算法的输入为训练集 $\{(x_j, y_j) \mid j = 1, 2, \dots, N\}$,包含 N 个训练样本的集合, y 为类标签。Adaboost 算法中,简单分类器 C_i 依赖于它的错误率 ϵ_i 。

给出经典的 Adaboost 算法(二值分类)的伪代码描述。

令 $\{(x_j, y_j) \mid j = 1, 2, \dots, N\}$ 表示包含 N 个训练样本的集合, y_j 为类别标签。 $x_i \in X$, X 表示领域或实例空间。 $y_j \in Y$ 。简单分类算法接受的样本是从分布为 P 的 $x \times y$ 上随机选择出来的。Adaboost 算法描述如下:

1: $w = \{w_j = 1/N \mid j = 1, 2, \dots, N\}$ 。{初始化样本权值。}

2: for $i = 1$ to T do

3: 根据 w ,通过对 D 进行抽样(有放回)产生训练集 D_i 。在 D_i 上训练简单分类器 C_i ,用 C_i 对原训练集 D 中的所有样本分类。

4: 错误率 $\epsilon_i = \frac{1}{N} [\sum_j w_j \delta(C_i(x_j) \neq y_j)]$, if $\epsilon_i > 0.5$ break。

5: end if

6: $\alpha_i = \frac{1}{2} \ln \frac{1 - \epsilon_i}{\epsilon_i}$

7: 根据公式

$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \times \begin{cases} e^{-\alpha_i} & \text{如果 } C_j(x_i) = y_i \\ e^{\alpha_i} & \text{如果 } C_j(x_i) \neq y_i \end{cases}$$

更新每个样本的权值。

8: end for

9: $C^*(x) = \arg_{y \in Y} \max \sum_{j=1}^T \alpha_j \delta(C_j(x) = y)$

T 为提升次数,错误率为 ϵ_i , $C^*(x)$ 为算法输出的最终分类器。Adaboost 算法利用样本的权值来确定其训练集的抽样分布。开始时,所有样本被赋予相同的权值 $1/N$,将样本的权重信息看作一个概率分布,按照此概率分布在原始样本集中抽样生成训练集 D_i ,然后使用这些训练集训练简单分类器 C_i ,用 C_i 对原数据

集中的所有样本进行分类。每一轮提升结束时更新训练样本的权值,增加被错误分类的样本的权值,减少被正确分类的样本的权值。这样,权重较大的样本有更好的机会被选中。这迫使分类器在随后的迭代中关注那些很难分类的样本。整个过程如此迭代进行,直到满足结束条件为止。

在每轮迭代后,虽然那些易于分类样本的权值会被更新为小于 $1/N$ 的数值,但在接下来的样本抽取中,依然可能会被选中,从而会出现对那些易于分类样本的重复抽取现象,影响到抽样时所获得的训练集 D_i 的有效容量。为了避免对易于分类样本的重复抽样,增加训练集 D_i 的有效容量,对原始算法有以下的改进。

3 样本标记的 Adaboost 改进算法

样本集 $\{(x_j, y_j, s_j) \mid j = 1, 2, \dots, N\}$ 包含 N 个训练样本, y_j 为类别标签, s_j 为选择标签。改进后的算法,即样本标记的 Adaboost 算法伪代码描述如下:

1: $w = \{w_j = 1/N \mid j = 1, 2, \dots, N\}, s_j = T$ 。{初始化样本权值,置可选标签为 T }。

2: for $i = 1$ to T do

3: 根据 w ,通过对 D 中满足 $s_j = T$ 的样本进行抽样(有放回)产生训练集 D_i 。在 D_i 上训练简单分类器 C_i ,用 C_i 对原训练集 D 中的所有样本分类。

4: 错误率 $\epsilon_i = \frac{1}{N} [\sum_j w_j \delta(C_i(x_j) \neq y_j)]$, if $\epsilon_i > 0.5$ break。

5: end if

6: $\alpha_i = \frac{1}{2} \ln \frac{1 - \epsilon_i}{\epsilon_i}$

7: 根据公式

$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \times \begin{cases} e^{-\alpha_i}, & \text{如果 } C_j(x_i) = y_i \\ e^{\alpha_i}, & \text{如果 } C_j(x_i) \neq y_i \end{cases}$$

更新每个样本权值,对于 $w_j < 1/N$ 的样本,置 $s_j = F$ 。

8: end for

9: $C^*(x) = \arg_{y \in Y} \max \sum_{j=1}^T \alpha_j \delta(C_j(x) = y)$

样本可选标签 s_j 标志样本在抽样时是否可被选取的状态。初始化时,样本集中所有样本的可选标签 s_j 均被置为 T (可选),即第一次对样本集进行抽样时,所有样本均可能被选中。

样本在提升(Boost)过程中,权值可能会被更新,若一旦某一样本的权值被更新为小于 $1/N$ 的数值,则其可选标志位 s_j 被相应的置为 F ,这样,该样本在以后的迭代过程中不可被选取作为训练集中的样本。在接下来的迭代过程的样本抽样中,只能取那些可选标志 s_j 为 T 的样本。相对于经典 Adaboost 算法迭代过程中

的抽样,改进后的算法提高了样本抽样的有效容量,避免了对于那些权值小于 $1/N$ 样本的重复抽取,这就使得改进后的算法能够快速地收敛并获得良好的精度。

4 实验

数据集的选取:采用 UCI 数据挖掘样例库 Repository of machine learning databases 中的数据集 Ionosphere 作为实验数据集,数据集包含 350 条数据,35 个属性^[5]。先对数据集中样本的类别标签(即最后一个属性)作相应预处理,使其符合二值分类的类别要求 $(-1, +1)$ 。添加可选标志位 s_j 作为附加属性,并置其为 T 。选取决策树作为简单分类器,Matlab 为实验平台,分别对样本使用经典的 Adaboost 算法与改进的 Adaboost 算法进行分类实验,并把分类结果作比较。实验结果如图 1 所示。

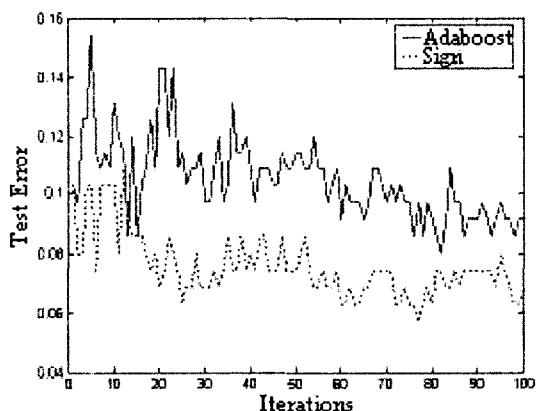


图 1 实验结果比较

图中纵向为算法对实验样本分类的错误率,横向为算法的迭代次数,曲线 Adaboost 为经典 Adaboost 算法在对实验数据集分类后的错误率-迭代次数曲线,曲线 Sign 是改进的 Adaboost 算法对实验数据集分类后的错误率-迭代次数曲线。

如图所示,与经典的 Adaboost 算法相比较,在迭代次数较少时,二者的错误率-迭代次数曲线区分并不明显,但随着迭代次数的增加,由于改进的算法在提升过程中,不断地对训练集中那些样本因权值更新后权重小于均值 $1/N$ 的样本置可选标志 s_j 为 F ,使得改进后的算法能更快速地关注那些难以区分的样本。可见,在迭代次数适当增加后,样本标记的 Adaboost 算法与经典的 Adaboost 算法相比较,获得了更快的收敛速度和更好的精度。

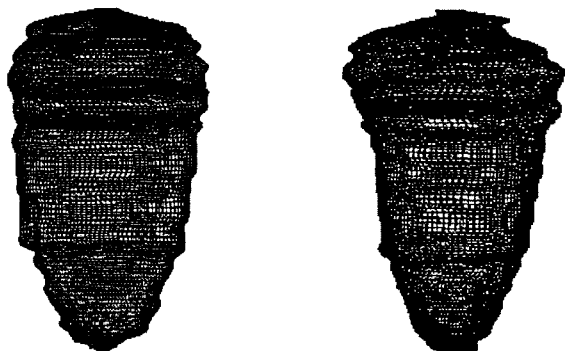
5 结束语

提出了一种对经典的 Adaboost 算法的改进算法,

(下转第 115 页)

$\dots, n(j))$, 定义误差限

$$E = \varepsilon \cdot \sum_{i=2}^{n(r)} |Q_{ij} - Q_{i-1j}|$$



(a) 正面图

(b) 侧面图

图 2 牙齿的三维实体图

曲线拟合时得到对应的参数值 \tilde{u}_{ij} , 对重新采样后的轮廓线数据点插值得到新的插值曲线 $c_j(u)$, 如果控制点 Q_{ij} 与 $c_j(\tilde{u}_{ij})$ 在误差限范围内, 就认为满足逼近要求, 对第 7 层 CT 图片, 当 $\varepsilon = 0.005$ 时, 误差限 $E = 0.5585$, 重新采样后的数据点满足误差要求 (如图 3 所示)。图 4 给出了对 39 层轮廓线进行误差分析的结果图, 通过分析, 满足逼近条件。

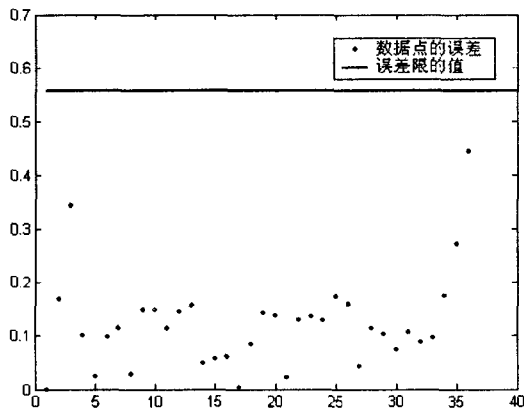


图 3 第 7 层数据点的误差图

文中算法是基于单轮廓线来进行, 能否在多轮廓

线重构中实现需要进一步的理论研究。

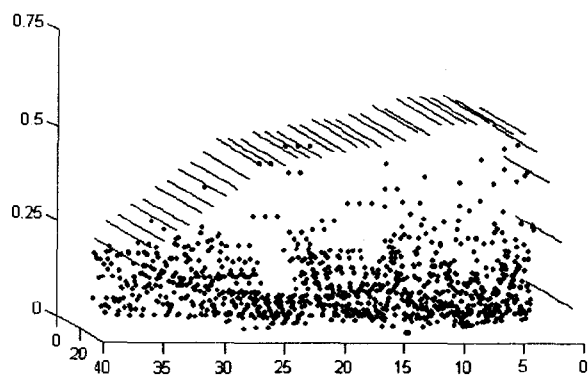


图 4 39 层数据点的误差图

参考文献:

- [1] 施法中. 计算机辅助几何设计与非均匀有理 B 样条 [M]. 北京: 高等教育出版社, 2001: 254-294.
- [2] 曲学军, 宁涛, 席平. 逆向工程中平面轮廓线数据的 B 样条曲面拟合 [J]. 计算机工程, 2004, 30(10): 14-19.
- [3] 李文杰. 利用三次 B 样条逼近断层轮廓构造封闭曲面 [J]. 福建电脑, 2005(7): 58-60.
- [4] 杨扬. 基于 CT 数据的三维曲面造型及应用 [D]. 西安: 西北大学, 2001.
- [5] 张毓晋. 图像处理 [M]. 北京: 清华大学出版社, 1999: 179-187.
- [6] 祁伟丽, 秦新强, 王溪. 基于二维平行轮廓线重建三维表面的算法研究 [C] // CIS&CSSS. 全国第 18 届计算机技术与应用学术会议论文集. 合肥: 中国科学技术大学出版社, 2007: 985-989.
- [7] 叶铭, 于力牛, 王成焱. 目标组织轮廓的三次非均匀 B 样条逼近 [J]. 上海交通大学学报, 2003, 37(5): 729-732.
- [8] Piget L A, Tiller W. Parametrization for surface fitting in reverse engineering [J]. Computer Aided Design, 2001, 33: 593-603.
- [9] Fisher R B. Applying knowledge to reverse engineering problems [J]. Computer Aided Design, 2004, 36: 501-510.

(上接第 111 页)

并通过实验证明, 该方法通过添加对训练样本的可选标记位, 使得算法在时间复杂度适当增加的同时, 更快速地在提升过程中关注那些难以分类的样本, 提高了算法的精度。

参考文献:

- [1] 李斌, 王紫石, 汪卫. AdaBoost 算法的一种改进方法 [J]. 小型微型计算机系统, 2004, 25(5): 869-871.
- [2] 郭红刚, 方敏. AdaBoost 方法在入侵检测技术上的应用

[J]. 计算机应用, 2005, 25(1): 144-146.

- [3] 杨宏晖, 孙进才, 牛奕龙. 支持向量机集成和特征选择联合算法 [J]. 声学技术, 2006, 25(4): 337-340.
- [4] Nock R, Nielsen F. A Real generalization of discrete AdaBoost [J]. Artificial Intelligence, 2007, 171: 25-41.
- [5] Collins M, Schapire R E. Logistic Regression, AdaBoost and Bregman Distances [J]. Machine Learning, 2002, 48: 253-285.