连续型 Adaboost 算法研究

严 超 王元庆
(南京大学电子科学与工程系 南京 210093)

摘 要 现阶段的人工智能与模式识别工作中,连续型 Adaboost 算法以其良好的识别率和极快的识别速度得到了越来越多的应用。鉴于此,认真研究了连续型 Adaboost 算法的理论基础,细致分析了基于连续型 Adaboost 算法的分类器的训练流程,对算法中涉及到的数学量之间的关系进行了探讨,对算法中涉及到的数学过程进行了定量推导,对训练过程中出现的问题的成因进行了定性分析,最后对如何提高连续型 Adaboost 算法的性能提出了若干建议。

关键词 连续型 Adaboost 算法,PCA 模型,归一化因子,检测率,过学习现象

中图法分类号 TP301.6 **文献标识码** A

Research of the Real Adaboost Algorithm

YAN Chao WANG Yuan-qing
(Department of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China)

Abstract In the current artificial intelligence and pattern recognition, Real Adaboost Algorithm, as for high accuracy rate and very fast speed, has been used more widely. As a result, we researched the theoretical basis of the Real Adaboost Algorithm conscientiously and analyzed the training procedures of classifiers based on the Real Adaboost Algorithm meticulously. In this course, we probed into the relationship between the mathematical variables involved in the algorithm; deduced the mathematical process involved in the algorithm quantitatively, and analyzed the reasons of problems appearing in training procedures qualitatively. At last, in order to improve the Real Adaboost Algorithm, we brought up several suggestions.

Keywords Real Adaboost algorithm, PCA model, Normalization factor, Testing rate, Excessive learning

如何快速通过训练方法获得准确率高的分类器是目前基于学习的模式识别领域的热点问题。针对这一问题,近几年出现了一大批分析角度不同、处理方法各异的优秀算法。其中,由 Freund 和 Schapire 于 1990 年提出的 Boosting 算法[1]以训练速度快、所得分类器准确率高而获得广泛的研究和应用。Boosting 的原意为提升、加强,其算法的中心思想是通过整合和训练,将弱分类器提升为强分类器。所谓弱分类器,是指该分类器的算法正确率刚刚超过 50%,即略好于随机猜想;所谓强分类器,是指其算法正确率远远高于 50%。另外,Boosting 算法还具有不需要任何训练样本的先验知识的优点。但是,因为 Boosting 算法中样本的权重无法"自适应"地调整,所以其对分类困难的样本的学习能力有限。

Adaboost 算法的全称是 Adaptive Boosting 算法,是由 Freund 和 Schapire 于 1995 年[2]提出的一种能够"自适应"的 Boosting 算法。Adaboost 算法能够自适应地调节训练样本权重的分布,不断挑选出当前样本权重分布下最佳的弱分类器,并整合所有得到的弱分类器,让它们按照一定的权重投票,组成一个强分类器。其中,样本权重的自适应调节,使得算法将检测重点"聚焦"于难以辨识的对象,提高了算法对分类困难的样本的学习能力,是对 Boosting 算法的一种优化。

本文对连续型 Adaboost 算法流程进行了细致分析,对算法中所用到的数学量和计算方法进行了定量推导,对算法涉及到的数学量的作用、数学量间的关系和训练过程中出现的某些问题的成因进行了定性分析。本文对连续型 Adaboost 算法的研究集中在:①归一化因子 Z 的作用及其选取条件的数学推导;平滑因子 ϵ 的作用推证。②算法检测率与算法误检率的相互关系及其影响因素的公式说明和定性分析。③连续型输出值的意义说明;过学习现象与偏见现象的成因。④提高算法性能的若干建议。

1 PAC 学习模型

可学习理论分为统计学习理论和计算学习理论两大部分[3]。作为 Adaboost 算法理论基础的 PAC(Probably Approximately Correct)学习模型属于计算学习理论的经典模型之一。PAC 学习模型在大量实验的基础上,对涉及到的一些量进行假设估计,在所使用的样本集样本数目增大、样本结构变复杂的情况下,依据概率理论研究上述的估计量能否一致收敛到未知真值的问题。

到稿日期:2009-10-25 返修日期:2009-12-29 本文受国家自然科学基金重点项目(60832003),上海大学新型显示技术及应用集成教育部重点实验室(P200902)资助。

严 超(1986-),男,博士,主要研究方向为模式识别、立体显示技术、立体视觉信息处理,E-mail:yanchao3756@yahoo.cn;王元庆(1964-),男,教授,博士生导师,主要研究方向为模式识别与人工智能、光学成像、立体视觉信息处理、显微图像处理、超模式图像处理。

PAC模型是由Valiant于1984年首先提出来的[4]。Valiant认为"学习"是当明显清晰的过程或模式不存在时仍能获取知识的一种"过程",并给出了一个从计算角度来获得这种"过程"的方法,这种方法包括:(1)适当信息收集机制的选择;(2)学习的协定;(3)对能在合理步骤内完成学习的概念的分类[5]。PAC学习的实质就是在样本训练的基础上,参照算法正确性和算法复杂度,构造最有效的算法使算法的输出以概率接近未知的目标概念。

PAC学习模型的数学描述如下。

首先给出:

1. X 为样本空间,包含所有可以用于学习的样本集合;

2. C 为概念空间,包含所有可以选取的目标概念 T;

3. V 为分类集合,其值为目标概念的所有分类 $\{V_1,V_2,\cdots,V_k\}$。最简单的情况为二值,$V=\{0,1\}$;

4. H 为假设空间,包含算法所输出的所有假设 $H_m(T,x)$。

学习器 L 的目的是找到目标概念的一个假设,使其能对每个样本进行分类。按照某种固定的(可能未知的)分布 $P(x)$ 独立抽取样本 $x_1,x_2,\cdots,x_m$,L 返回 $H_t(x_t)$ 的值:$H_t(x_t)\in V$ 为 T 的指示函数,表示 L 对 x_t 的分类。于是可以获得一组数据:

$$[(x_1,h_t(x_1)),\cdots,(x_m,h_t(x_m))]\in[X,V]^m$$

构造适当的算法 $\{A_m\}$,A_m 为到概念空间的映射。$A_m:[X,V]^m\to C$,并定义:

$$H_m(T,x)=A_m((x_1,h_T(x_1)),\cdots,(x_m,h_T(x_m)))$$

则 $H_m(T,x)$ 就是目标概念 T 对样本 $x_1,x_2,\cdots,x_m$ 的一个假设。

我们希望能够找到一个对所有样本都正确的假设,在实际学习中,这是不可能的。如果学习器 L 最终将以 $(1-\delta)$ 的概率(δ 称为假设的置信度)输出一个假设 $h\in H$,而且随机样本被错误分类的概率小于假设错误率 ε,就认为这个假设为成功假设。如果学习器 L 只需要多项式 $p(m,1/\varepsilon,1/\delta)$ 个样本以及在多项式 $p(m,1/\varepsilon,1/\delta)$ 时间内就可以获得一个成功假设,那么称 L 为可PAC学习的。

2 连续型Adaboost算法介绍

Viola等人于2001年提出了一种基于Haar型特征,利用积分图进行运算的离散型Adaboost算法,并将其与Cascade结构结合,学习出基于离散型Adaboost算法的瀑布型人脸检测器。后来Schapire等人对离散型Adaboost算法进行改进,使得弱分类器的输出具有了连续的置信度,也使得算法的收敛速度得以加快,并在此基础上学习出了基于连续型Adaboost算法的瀑布型人脸检测器。这种基于连续型Adaboost算法和Cascade结构的人脸检测器是目前检测正确率最高的人脸检测系统之一,同时其检测速度也远远快于几乎所有基于其他算法的人脸检测器。

连续型Adaboost算法训练流程描述如下:

(1)给定训练样本集合 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,弱分类器空间 H,其中 $x\in X$ 为样本向量,$y=\pm1$ 为类别标签,n 为样本总数。初始化样本概率分布 $D_t(i)=1/n$,$i=1,2,\cdots,n$。

(2)For $t=1,2,\cdots,T$(T 为需要得到的特征数目)

①对 H 中的每个弱分类器 h 做如下操作:

1)对样本空间 X 进行划分,得到:$x_1,x_2,\cdots,x_m$;

2)在训练样本的概率分布 D_t 下,计算:

$$W_k^i=P(x_i\in X_j,y_i=k)=\sum_{\substack{x_i\in x\\ y_i=k}}D_t(i),k=\pm1 \tag{1}$$

3)设置弱分类器在这个划分上的输出:

$$\forall x\in X_j,h(x)=\frac{1}{2}\ln\left(\frac{W_{+1}^j+\varepsilon}{W_{-1}^j+\varepsilon}\right),j=1,2,\cdots,m \tag{2}$$

式中,ε 为一小正常数。

4)计算归一化因子:

$$Z=2\sum_j\sqrt{W_{+1}^jW_{-1}^j} \tag{3}$$

②在弱分类器空间中选择一个 h_t,使得 Z 最小:

$$Z_t=\min Z,h\in H$$

$$h_t=\operatorname{argmin}\ Z,h\in H$$

③更新训练样本概率分布:

$$D_{t+1}(i)=D_t(i)\frac{\exp[-y_ih_t(x_i)]}{Z_t} \tag{4}$$

式中,Z_t 为归一化因子,使得 D_{t+1} 为一个概率分布。

(3)最终强分类器 H 为:

$$H(x)=\operatorname{sign}[\sum_{t=1}^{T}h_t(x)-b] \tag{5}$$

式中,b 为手动设定的阈值,默认为0。

类似地可以定义 H 的致信度:

$$conf_H(x)=|\sum_t h_t(x)-b| \tag{6}$$

3 连续型Adaboost算法探究

3.1 归一化因子 Z

从上述训练流程可知,归一化因子 Z 是弱分类器空间中特征选取的唯一参照标准。因此,首先对 Z_t 作为归一化因子进行定量论证并对 Z 的最小化选取进行定性分析。从训练样本概率分布的更新表达式(1)开始可以推得:

$$\begin{aligned}&\sum_i D_t(i)\exp[-y_ih_t(x_i)]\\&=\sum_j\{\sum_{y=1}D_t(i)\exp[-h_t(x_i)]+\sum_{y=-1}D_t(i)\exp[h_t(x_i)]\}\\&=\sum_j\{[\sum_{y=1}D_t(i)]\cdot\sqrt{\frac{W_{-1}^j}{W_{+1}^j}}+[\sum_{y=-1}D_t(i)]\cdot\sqrt{\frac{W_{+1}^j}{W_{-1}^j}}\\&=\sum_j(W_{+1}^j\cdot\sqrt{\frac{W_{-1}^j}{W_{+1}^j}}+W_{-1}^j\cdot\sqrt{\frac{W_{+1}^j}{W_{-1}^j}})\\&=2\sum_j\sqrt{W_{+1}^j\cdot W_{-1}^j}=Z_t\end{aligned}$$

完成了对 Z_t 作为归一化因子的定量论证之后,我们通过对归一化因子 Z 的计算式进行分析可以得到:在样本权重归一化的前提下,同一划分区域内正负样本的权重相差越大,Z 的取值越小,这时,正负样本也越容易区分,分类器的性能也越好;同理,同一划分区域内正负样本的权重相差越小,Z 的取值越大,分类器的性能也越差。所以,选取 Z 值最小的弱分类器为最佳弱分类器是完全正确的。

另外,弱分类器的输出公式(2)中的小正常数 ε 称为平滑因子,其作用是在一定程度上将弱分类器的输出限定在一定范围内。假设弱分类器 $h(x)$ 的表达式(2)中不含有平滑因子 ε。在一个弱分类器中,当某一划分区域内正负样本比重相差悬殊时,落在该区域内的样本的输出值将是绝对值意义上的大数。如果这个弱分类器是组成强分类器的一员,则由强分类器 $H(x)$ 的表达式(5)可知,很可能会因为这个弱分类器的

输出值过大而造成对该种特征的过学习。平滑因子的加入则可以有效地避免弱分类器的输出值过大，从而也可有效地避免过学习现象的产生。

3.2 检测率与误检率

前面已经提到，连续型 Adaboost 算法属于计算学习理论范畴，其数学依据主要为概率理论。从训练样本的概率计算公式(1)和弱分类器的输出计算公式(2)可以看出，落在同一划分区间内的样本的输出值完全相同，输出值的正负取决于该划分区域内样本权重的概率分布。

首先，从微观层面研究弱分类器的检测率与误检率。由概率论的知识可知，在训练样本为大量的情况下，当某一划分区域内正样本的权重总和远大于负样本的权重总和时，我们完全有信心将该区域的输出设为正值；反之，当某一划分区域内负样本的权重总和远大于正样本的权重总和时，我们完全有信心将该区域的输出设为负值。但是，当某一划分区域内正样本的权重总和与负样本的权重总和相差不大时，则在设置该区域的输出时会显得信心不足：如果某一划分区域内负样本的权重总和略大于正样本的权重总和，将该区域的输出置负，那么很可能导致相关弱分类器检测率的下降；如果某一划分区域内正样本的权重总和略大于负样本的权重总和，就将该区域的输出置正，那么很可能导致相关弱分类器误检率的上升。由此可见，在训练样本为大量的情况下，正样本权重总和与负样本权重总和相差不多的区域，其输出最有可能对弱分类器的检测率及误检率造成影响。

其次，从宏观层面研究强分类器的检测率与误检率。对照强分类器的输出公式(5)可知，阈值 b 的选取对强分类器的检测率和误检率影响巨大：阈值 b 的理论取值为 0，这是建立在各弱分类器的阈值都为 0 的基础上的；根据上段所述，当阈值 b 变大时，输出为小正值的负样本会被判定为负，从而降低了分类器的误检率，但是同时输出为小正值的正样本也会被判定为负，从而也会降低分类器的检测率，在样本数量为大量、样本分布比较离散的情况下，检测率的下降甚至更为急剧；反之，当阈值 b 变小时，输出为小负值的正样本会被判定为正，从而提高了分类器的检测率，但是同时输出为小负值的负样本也会被判定为正，从而也会提高分类器的误检率，在样本数量为大量、样本分布比较离散的情况下，误检率的上升甚至更为急剧。因此，我们会经常遇到一些没有漏检(检测率 100%)，误检却很多(误检率同样很高)的人脸检测器，这正是为了追求高检测率，调低阈值 b，但同时也牺牲了误检率的结果。

3.3 连续型输出值与过学习现象

1996 年，Freund 和 Schapire 就关于分类器的训练问题提出了离散型 Adaboost 算法[6]，随后，Schapire 等人就提出了连续型 Adaboost 算法[7]。这主要是因为相对于离散型 Adaboost 算法，连续型 Adaboost 算法能够更加精确地刻画分类器的边界并可以取得连续的置信度。连续型 Adaboost 算法是基于灰度的人脸检测算法，任何区域(包括人脸)基于灰度所得到的特征准确地说应当是一个或几个特征区间(人种不同，基于灰度的人脸特征空间可能不同)，而不是某一个精确的特征数值。连续型 Adaboost 算法正是依据这一点和概率理论确定其分类器的训练方法的。

对比离散型 Adaboost 算法的训练流程[5]和连续型 Adaboost 算法的训练流程可知：1)连续型 Adaboost 算法关于弱分类器的训练是将样本的特征值划分为若干区间，然后各个区间按照两类样本的着落概率确定输出；离散型 Adaboost 算法却是将样本的特征值仅仅划分为两个区间。特征值划分区间的数目越多，算法对不同检测目标输出值的刻画也越具体，算法检测率自然也就越高，同时，算法的适应域也越广泛，鲁棒性也越好。2)连续型 Adaboost 算法每个弱分类器的输出都是连续实值，这些实值对 0 值(理论阈值)的偏离度本身就代表着弱分类器对既定目标的分类置信度；而离散型 Adaboost 算法中的弱分类器则仅有两个输出值：0 或者 1，只能进行二值断言，无法对自身的分类判断提供置信度依据。因此，在弱分类器整合为强分类器的过程中，相对于二值断言的离散型 Adaboost 算法，连续型 Adaboost 算法通过组合更为具体、包含信息量更为丰富的连续实值输出，无疑更能够提供正确的分类判断。

过学习现象是指，在训练过程中，由于分类器对某类特征的过分学习而造成的该种特征权重过大从而影响算法准确率的一种样本权重扭曲分布的现象。相反，偏见现象是指在训练过程中，由于分类器对某类特征的学习不足而造成的该种特征权重过小从而影响算法准确率的另一种样本权重扭曲分布的现象。

J. R. Quinlan 对 Boosting 和 Bagging 两种机器学习的优化算法曾做实验进行比较[8]，在一个叫做 Iris 的数据集上，原本权重比例基本相等的 3 个预测目标类 setosa，versicolor 和 virginica 在循环迭代 5 次后，权重分布出现了急剧的扭曲，表现为 setosa 类上的样本权重仅占不到全部样本权重的 3%，而 versicolor 类上的样本权重所占比例却接近全部样本权重的 80%。从训练样本的权重更新式(4)可以看出，难于识别样本的权重可能会随着样本权重的迭代更新而不断增加，从而在整个样本集上实现扩张，这样，训练得到的分类器很可能会对某一目标类产生过度学习现象。同时，容易区分样本的权重可能会随着样本权重的迭代更新而不断减小，一些正确的分类规则因此会被破坏甚至丢弃，从而产生对分类规则的偏见[9]。

4 算法性能提高建议

1. 进一步细化查找表，增加步骤(1)中的样本分区数目。所谓查找表，即同一特征的不同分区与不同输出值之间的对应关系。从 3.3 节两种 Adaboost 算法特征值区间划分的对比中我们得到：特征值划分区间的数目越多，算法对不同检测目标输出值的刻画也越具体，算法的性能越好。比如，将一个原来输出值为 5 的特征区间进一步划分为输出值为 2,4,8 的 3 个特征空间，显然后者比前者更能提供细致准确的输出值和置信度。当然，查找表的细化工作必须建立在训练样本数目为大量的基础上。当训练样本数目有限时，依据样本在分区上的着落概率进行分类预测的随机性大大增强，此时，过多的分区反而会加大这种随机性。因此，当训练样本数目有限时，应当相应地限制步骤(1)中的样本分区数目。

2. 确定样本权重上限，防止样本权重的过度扩张或减小。通过 3.3 节中对过学习现象的研究可以看出，为防止过学习现象及偏见现象，必须对训练样本的权重加以限制。限制样

(下转第 248 页)

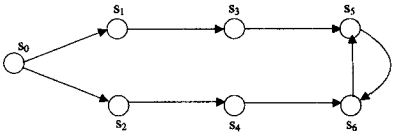

图 5 Model

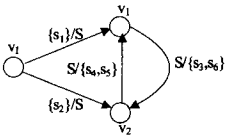

图 6 断言图

表 1 GSTE(G,post)

	V1,V1	V1,V2	V1,V2	V2,V1
0	{s1}	T	{s2}	T
1	{s1}	{s3}	{s2}	{s4}
2	{s1}	{s3,s6}	{s2}	{s4,s5}
4	{s1}	{s3,s6}	{s2}	{s4,s5}

其中，*err* 为 V1V2。

首先求 s'：

$s'=\mathrm{sim}(err)-\mathrm{cons}(\mathrm{err})=\{s6\}$

应用 2.3 节提出的算法来解决，详细过程如表 2 所列。

表 2 CE(Tree,pre,err)

	V1,V1	V1,V2	V1,V2	V2,V1	Link	s'
0	{s1}	{s3,s6}	{s2}	{s4,s5}	V1V2	{s6}
1	{s1}	{s3,s6}	{s2}	{s4,s5}	V1V2→V2V1	{s4,s5}
2	{s1}	{s3,s6}	{s2}	{s4}	V1V2→V2V1→V1V2	

通过表 2 可以看出，此方法可以很好地解决反例查找问题，可以方便地查找到反例。

结束语 本文在 GSTE 的理论背景下，提出了一种寻找反例算法，算法通过在模拟过程中记录下每个边 e 的每一个 sim(e)，无论计算过程中怎么抽象，都可以找到产生错误的前一条边。并将采用符号变量的路径节点放入 hash 表中，每次都检索 hash 表中是否有此节点，如果有就对其进行路径选择，从而找到正确的反例路径。

如果此算法引入启发式的算法，通过它来解决处理边的循环和分裂，反例的查找会更高效。

参 考 文 献

[1] ITRS2003[OL]. http://public.itrs.net/

[2] 韩俊刚，杜惠敏. 数字硬件的形式验证[M]. 北京：北京大学出版社，2001

[3] 刘楠，朱文也，祝跃飞，等. 基于树语言逼近的安全协议形式化分析[J]. 计算机科学，2010(1)

[4] Jin Yang, Seger C-J H. Generalized Symbolic Trajectory Evaluation[C]// Proceedings of 2001 IEEE International Conference on Computer Design. 2001:360-365

[5] Jin Yang, Seger C-J H. Introduction to generalized symbolic trajectory evaluation[J]. IEEE transactions on very large scale integration(VLSI) systems, 2003, 11(3)

[6] Matthew B, Dwyer Corina S, et al. Finding feasible counter-examples when model checking Java programs[C]// Tools and Algorithms for the Construction and Analysis of Systems (TACAS), volume 2031 of Lecture Notes in Computer Science. Springer-Verlag, 2001

[7] Ball T, Rajamani S K. Checking temporal proper-ties of software with boolean programs[C]// Workshop on Advances in Verication(with CAV 2000). 2000

[8] Chen Yan, He Yujing, Xie Fei, et al. Automatic Abstraction Refinement for Generalized Symbolic Trajectory Evaluation[C]// Proceedings-Formal Methods in Computer Aided Design, FMCAD 2007. 2007:111-118

(上接第 211 页)

本权重分为两种情况：当训练样本明显地分为若干目标类时，应当首先确定这些目标类的权重比例，然后把每个样本的权重调整限定在类内完成。比如 3.3 节中涉及到的 Iris 样本集，可首先将其中包含的 3 个目标类的权重比例设为 $a:b:c$ ($a+b+c=1$)，并使得这个比例在每轮循环中都得以保持，然后在每一轮的循环中，各预测目标类按照该权重比例，在目标类内部更新样本权重，进行标准化[9]；当训练样本不能够明显地分为若干目标类时，可以根据训练样本的数量设置单个样本的权重上限和权重下限，将单个样本的权重限定在一定区间内，避免最终得到的分类器对样本产生过学习或偏见现象。

3. 尽量增加训练样本数目，样本选取尽量离散化。从上述对连续型 Adaboost 算法的探究可知，算法最重要的基础是概率理论，由训练结果得到检测结果的过程包含一个先验概率问题，即由训练样本在各个分区上的着落概率决定检测目标值的分类结果。因此为避免随机性，应当尽量增加训练样本的数目，以满足算法的概率理论基础。同时，为扩大所得分类器的适用范围，提高所得分类系统的鲁棒性，样本的选取应当尽量离散化，如人脸样本库中的样本应当来源于不同人种、不同年龄段，带有不同表情，在不同的光照条件下获得。只有样本离散化达到一定程度同时样本数目达到一定水平的样本库所训练出来的分类器，其性能才有一定保证。

参 考 文 献

[1] Schapire R E. The strength of weak learnablhty[J]. Machine Learning, 1990, 5(2):197-227

[2] Freund Y. Boosting a weak learning algorithm by majority[J]. Information and computation, 1995, 141(2):256-285

[3] Song Chunlei, Wang Long. Learning Theory and Robust Control [J]. Journal of Control Theory and Application, 2000, 17(5): 633-636

[4] Valiant L G. A Theory of the Learnable[J]. Communications of the ACM, 1984, 27(11):1134-1142

[5] Zhao Nan. Face Detection Based on Adaboost[D]. Peiking University, 2005

[6] Freund Y. Schapire R E. Experiments with a New Boosting Algorithm[C]// Proc. the 13th Conf. Machine Learning. San Francisco: M organ Kaufmann, 1996:148-156

[7] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37 (3):297-336

[8] Quinlan J R. Bagging, bosting. and C4. 5[C]// Proceedings of the Thirteenth National Conference on Artificial Intelligence. 1996:725-730

[9] Li Bin. Enhancing Method for Adaboost[J]. Mini-Micro Systems, 2004, 25(5)