第46卷第3期 中国科学技术大学学报 Vol. 46, No. 3
2016年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2016

文章编号:0253-2778(2016)3-0222-09

Boosting 算法理论与应用研究

张文生,于廷照

(中科院自动化所, 北京 100190)

摘要:作为机器学习领域最经典算法之一,Boosting 是一种学习算法,并广泛应用于机器学习与模式识别各领域. Boosting 的理论研究分为可学习理论和统计学两个角度. Boosting 最初从弱可学习理论角度阐明了由弱到强的提升算法,从理论上证明了一组优于随机猜测的弱学习器通过集成可提升为在训练集上任意精度的强学习器. 从统计学的角度看,Boosting 是一种叠加模型,理论上二者的等价性已经证明. 本文首先从可学习的角度出发,回顾了 Boosting 算法弱可学习理论,并提出面临的问题及挑战,包括对高维数据的有效性及 Margin 理论;然后阐述了 Boosting 算法理论研究分支,并详细回顾了当前最为流行的多种经典 Boosting 算法及在 Boosting 理论框架下的新应用;最后探讨了 Boosting 算法的未来研究趋势.

关键词:Boosting;弱可学习理论;Margin 理论;集成学习;AdaBoost

中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.03.007

引用格式:ZHANG Wensheng, YU Tingzhao. Research on Boosting theory and its applications[J]. Journal of University of Science and Technology of China, 2016,46(3):222-230.
张文生,于廷照. Boosting 算法理论与应用研究[J]. 中国科学技术大学学报,2016,46(3):222-230.

Research on Boosting theory and its applications

ZHANG Wensheng, YU Tingzhao

(*Institute of Automation, Chinese Academy of Science, Beijing* 100190, *China*)

Abstract: Boosting is one of the most popular ensemble algorithms in machine learning, and it has been widely used in machine learning and pattern recognition. There are mainly two frameworks of Boosting, learnable theory and statistical theory. Boosting was first proposed from the theory of weak learnability which illustrates the theory of boosting a group of weak learners into a strong learner. After a finite number of iterations, the combination of weak learners could be boosted into any accuracy on the training set, and the only requirement for a weak learner is that the accuracy be slightly better than a random guess. From the statistical point of view, Boosting is an additive model, and the equivalence between these two models has already been proved. The theory of weak learnability is reviewed from the PAC perspective, and the challenges Boosting may face are presented, includeing effectiveness for high dimension data and the Margin theory. Then, various Boosting algorithms are discussed from the above two viewpoints and their new applications with Boosting framework. Finally, the future of Boosting is discussed.

Key words: Boosting; weak learnability; margin theory; ensemble learning; AdaBoost

收稿日期:2015-09-12;**修回日期**:2015-12-29

作者简介:张文生(通讯作者),男,1966 年生,博士/研究员. 研究方向:人工智能,数据挖掘. E-mail: wensheng.zhang@ia.ac.cn.

0 引言

Boosting 思想源于 1988 年[1] 由 Kearns 和 Valiant 抛出的一个理论问题:能否由一组弱学习器“提升”为高效的强学习器. Schapire[2] 于 1990 年对其进行了证明. Schapire 指出,当弱学习器的准确率优于随机猜测时,可以将任意一组弱学习器“提升”为在训练集上具有任意精度的强学习器,与此同时提出了 Boosting 算法,但此时 Boosting 算法并不实用. 1995 年,Schapire 等提出了 Boosting 算法的实用算法 AdaBoost[3-6].

1998 年 Friedman 等[7] 提出从统计学的角度看待 Boosting,并提出了 LogitBoost[7] 和 GradientBoost[8],将 Boosting 算法看作叠加逻辑回归模型,并证明了二者的等价性. 自此,关于 Boosting 算法的研究分为两个方向[9],即以 AdaBoost 为代表的 WTSL(weak to strong learner)和以 GradientBoost 为代表的 CWGD(coordinate wise gradient descent).

Boosting 算法的应用分两种类型:第一种即根据 Boosting 算法框架设计新的 Boosting 算法,这方面的研究包括 AdaBoost[3,7]、GentleBoost[7]、RankBoost[10]、GradientBoost[8]、Sto GradientBoost[11]、LazyBoost[12]、MultiBoost[13]、BrownBoost[14]、AnyBoost[15]、LPBoost[16]、L2Boost[17]、SmoothBoost[18]、RegBoost[19]、DistBoost[20]、Modest Boost[21]、GiniBoost[22]、TotalBoost[23]、SparseBoost[24]、FilterBoost[25]、SoftBoost[26]、SavageBoost[27]、RobustBoost[28]、TwinBoost[29]、DirectBoost[30] 和 StructBoost[31] 等;第二种应用则是将现有 Boosting 算法运用到工程中. Schapire 等将 Boosting 算法运用到文本分类中[32],James 等将 Boosting 算法运用到音乐流派和艺术家预测[33],Li 等将 Boosting 算法运用到目标检测与识别[34],Liu 等将 Boosting 算法运用到人脸表情识别[35]. 其中最著名的是 Viola 等将 AdaBoost 运用到人脸检测[36],极大地提高了人脸检测的准确率. 最近,Saberian 等[37] 在 Viola 等的基础上,提出了 FCBoost.

关于 Boosting 算法已有一些综述性的文章[38-43],本文从弱可学习理论出发,首先揭示 Boosting 算法的理论来源;然后指出 Boosting 算法遇到的问题及挑战,包括 Boosting 算法对于高维数据的有效性[44] 及 Boosting 算法不容易出现过拟合的 Margin[45] 解释;其次回顾了多种经典 Boosting 算法及其在最新理论框架下的新运用 GNNB(gentle nearest neighbor boosting)[46]、CRBoosting(collaborative representation boosting)[47] 和 online Boosting[48];最后在展望 Boosting 未来及趋势.

1 PAC 可学习到弱可学习

Boosting 算法的理论基础是弱可学习理论(weak learnability),而弱可学习的概念源自 PAC 可学习(强可学习)[49]. 即如果存在样本复杂度 $m_{\mathcal{H}}:(0,1)^2 \to \mathbb{N}$ 以及学习算法 A 对于假设类 $\mathcal{H}$,分布 D 以及标号函数 f 成立. 当学习算法作用于分布 D 产生的. 由标记函数 f 标定的 $m \geqslant m_{\mathcal{H}}$ 个独立同分布的样本时,会返回一个假设 h 使得误差 $L_{(D,f)}(h) \leqslant \varepsilon$ 的概率至少为 $1-\delta$,那么称假设类 $\mathcal{H}$ 是 PAC 可学习的. 其中算法对于任意的 $\varepsilon,\delta \in (0,1)^2$,任意 $\mathscr{X}$ 上的分布 D 以及任意标号函数 $f:x \to \{\pm 1\}$ 可实现[50].

1988 年,Kearns 等提出是否可以通过降低准确率来减小计算复杂度[1],并使得一组低准确率的弱学习器可以提升为高效的强学习器. 这就引出了弱可学习的定义:如果存在函数 $m_{\mathcal{H}}:(0,1) \to \mathbb{N}$ 使得对任意的 $\delta \in (0,1)$,任意 $\mathscr{X}$ 上的分布 D 以及任意标签函数 $f:x \to \{\pm 1\}$,可实现假设对于 $\mathscr{X}$,D,f 成立,那么,当学习算法作用于分布 D 产生的、由标记函数 f 标定的 $m \geqslant m_{\mathcal{H}}$ 个独立同分布的样本时,会返回一个假设 h 使得 $L_{(D,f)}(h) \leqslant 1/2-\gamma$ 的概率至少为 $1-\delta$,就称学习算法 A 是假设类 $\mathcal{H}$ 的 γ(弱可学习器)[50].

1990 年,Schapire[2] 首先解决了此疑问,指出在满足一定条件(误差 $\varepsilon < 1/2$)下,一组弱学习器确可提升为对于训练集上任意精度的强学习器,并提出了 Boosting 算法. 1995 年 Freund 等[3] 提出 AdaBoost,由于其简单实用并且有理论保证收敛性[51],Boosting 算法被广泛接受,应用到机器学习等各领域[52-53],适用于大数据的 Boosting 算法也应运而生[54].

2 问题与挑战

2.1 高维数据的挑战及有效性

机器学习算法除了面临海量难以处理数据的困惑,更为艰巨的一个挑战来自大数据的高维特性,这

也就衍生了 BigData[2,55]. 降维提供了一个可行的解决方案,但其计算量较大;核方法有效地解决了高维及计算量大的问题,但需要一定的先验知识,即核函数的选取依据.

Boosting 算法可以有效的解决高维数据带来的挑战,2003 年,Buhlmann 等[17]指出,Boosting 算法对于学习高维学习器具有很大的优势,Buhlmann 给出的一个解释[44]是 Boosting 在学习过程中进行了特征选择(假设弱学习器选择决策树,每个特征对应一棵决策树,每学习一个弱学习器,即相当于选择了一种特征)并且学得的弱学习器自由度可变.

核方法在高维空间巧妙地避免了计算量复杂的问题;Boosting 算法则是通过在高维空间中学习一组简单且易于高效实现的弱学习器,通过弱学习器的加权线性组合获得强学习器解决高维空间的问题,计算复杂度为 $O(dm)$ [50].

2.2 Boosting 的 Margin 解释

当数据量过大时,一是计算复杂,二是样本数量与模型复杂度不对称造成的过拟合. 通常情况下 Boosting 算法不会出现过拟合[56-58]. 首先,Boosting 算法的泛化误差并不随着迭代次数的增加而增大;其次,当 Boosting 算法的训练误差接近 0 时,泛化误差仍然随着迭代次数的增加而减小,如图 1 所示.

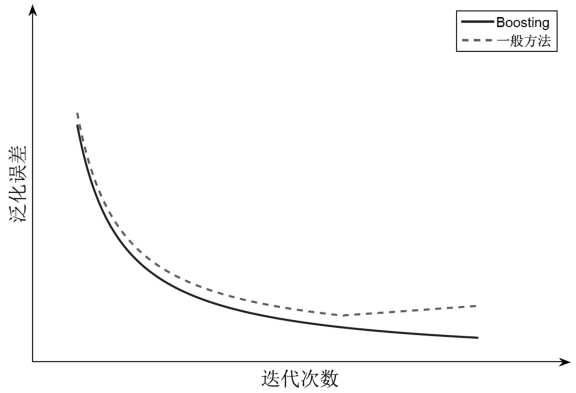

图 1 Boosting 与传统算法泛化误差随迭代次数的变化

Fig. 1 Changes of Boosting and the traditional algorithm generalization error with iteration

Brieman 最早给出没有过拟合的“Variance”[59]解释,并指出,Boosting 算法属于多数投票机制,多数投票机制的工作原理就是降低学习算法的“Variance”.

Schapire 提出质疑[60],并指出“Variance”只能解释多数投票机制中的 bagging[61]算法,而对于 Boosting 来说,“Variance”减小并不是必须的,并提出了 Boosting 没有过拟合的 Margin 解释,其中 Margin 定义为

$$\mathrm{Margin}(x_i,y_i)=\sum_{m=1}^{M}y_i\alpha_m h_m(x_i).$$

式中,x_i 表示第 i 个样本,y_i 表示 x_i 的标签,h_m 表示第 m 个弱学习器,α_m 表示 h_m 的权重,M 表示迭代次数,因此,$h_m(x_i)$ 就是第 t 个弱学习器对第 i 个样本的预测值,$\sum_{m=1}^{M}\alpha_m h_m(x_i)$ 则是样本 i 的最终预测值. 由定义可知,Margin 的值域为 $[-1,1]$,进一步,当 $\mathrm{Margin}<0$ 表示预测错误,而 $\mathrm{Margin}>0$ 表示预测正确,并且 Margin 幅值的大小表示预测正确错误的置信度. 事实上,Schapire 证明 Boosting 算法的泛化误差与样本 Margin 的分布、训练样本数量及弱学习器的复杂度(即 VC 维)有关[62].

1999 年,Breiman 根据 Margin 理论提出 ARC-GV 算法[63],这种算法可以最大化任意训练样本的最小 Margin,通过与 AdaBoost 的对比试验,证明 ARC-GV 算法确实能产生更大的 Margin,但 ARC-GV 算法的学习效果却不如 AdaBoost.

2006 年,Reyzin 等[62]指出,尽管 ARC-GV 方法确实有更大的 Margin,但这只是样本的最小 Margin,Boosting 算法的误差界与 Margin 的整体分布相关,而非最小 Margin.

近年来,ZHOU[64]和 GAO[65]提出,不能只考虑 Average Margin 而应该考虑 Margin 的“Variance”,并指出了泛化误差界与 Average Margin 和 Margin Variance 的关系.

3 研究现状分析

3.1 Boosting 算法研究分支

Boosting 算法理论的研究分为两个方向[9],第一个方向即以 AdaBoost 为代表的从弱可学习理论角度出发的 WTSL,第二个方向则是以 GradientBoost 为代表的从统计学角度出发的 CWGD. 1998 年,Friedman 等证明了二者的等价性[7]. WTSL 源自 Boosting 算法最初始的弱可学习理论,它的理论基础可以确保在有限的迭代步内学得有限的弱学习器,并且由此组成的强学习器的误差可以任意小. 其次,对于弱学习器的权重分配,WTSL 有完美的解决方案,并可保证其收敛性. 最后,损失函数、权重更新算法以及优化算法是 WTSL 框架的三个要点. WTSL 框架是一个迭代的

过程,除最基本的根据样本权重及标签学习弱学习器外,WTSL 包含三个相关,学习误差与样本权重相关、弱学习器权重与学习误差相关、样本权重与学习效果相关. 以 CWGD 框架的 Boosting 算法则不考虑算法的可学习性. 在 CWGD 框架下,凸损失函数的选择是设计新 Boosting 算法的关键,但凸函数的选择对 CWGD 又是一种局限,若要保证稀疏性或者平滑性,则选择最优损失函数依赖于先验知识. CWGD 框架对于弱学习器权重的分配没有明确的指示.

3.2 多种 Boosting 方法回顾

3.2.1　AdaBoost 算法

AdaBoost 算法是一个迭代的过程,包含弱学习器、样本权重、分类器权重三个基本点. 由于采样过程可以视作样本加权的直观体现,而样本加权是 AdaBoost 算法的突出特点之一,因此从某种意义上说,AdaBoost 是一种重采样方法. 一般认为 AdaBoost 分 Discrete 和 Real 两种类型,算法流程如算法 3.1 和 3.2 所示.

算法 3.1　Discrete AdaBoost

输入:训练集 X 及其对应标签 y,迭代次数 M

输出:强分类器 F

1 样本权重初始化为 $w_i=1/N, i=1,2,\ldots,N$,其中 N 为样本总数.

2 迭代获得强学习器:

FOR $m=1,2,\cdots,M$:

(1)在训练集上根据权重 w_i 学习获得弱学习器 $f_m\in\{\pm 1\}$;

(2)计算当前弱学习器误差及其权重

$$err_m=E_m[1_{(y\neq f_m(x))}]$$

$$c_m=\log\frac{1-err_m}{err_m};$$

(3)更新样本权重

$$w_i\leftarrow w_i\cdot\exp[c_m\cdot 1_{(y_i\neq f_m(x_i))}],\ i=1,2,\ldots,N$$

并归一化使得 $\sum_i w_i=1$;

END FOR

3 输出强分类器 $F(x)=\mathrm{sign}[\sum_{m=1}^{M}c_m f_m(x)]$

Real AdaBoost 和 Discrete AdaBoost 的不同在于弱学习器的输出不再是二值函数,而是表明置信度的概率值,其数值大小恰能反映弱学习器的权重.

算法 3.2　Real AdaBoost

输入:训练集 X 及其对应标签 y,迭代次数 M

输出:强分类器 F

1 样本权重初始化 $w_i=1/N, i=1,2,\ldots,N$,其中 N 为样本总数.

2 迭代获得强学习器:

FOR $m=1,2,\cdots,M$:

(1)在训练集上根据权重 w_i 学习获得概率分布 $p_m(x)=\hat{P}_w(y=1|x)\in[0,1]$;

(2)计算当前弱学习器输出

$$f_m(x)\leftarrow\frac{1}{2}\cdot\log\frac{p_m(x)}{1-p_m(x)};$$

(3)更新样本权重

$$w_i\leftarrow w_i\cdot\exp[-y_i f_m(x_i)],\ i=1,2,\ldots,N$$

并归一化使得 $\sum_i w_i=1$;

END FOR

3 输出强分类器 $F(x)=sign[\sum_{m=1}^{M}f_m(x)]$

3.2.2　LogitBoost 算法

Friedman 等在研究 Boosting 特性之后,于 1998 年首次提出 LogitBoost[7]. LogitBoost 是最早从统计学角度看待 Boosting 的算法,它利用牛顿法通过极大似然估计来拟合叠加模型. 算法流程如算法 3.3 所示.

算法 3.3　LogitBoost

输入:训练集 X 及其对应标签 y,迭代次数 M

输出:强分类器 F

1 样本权重初始化 $w_i=1/N, i=1,2,\ldots,N$,其中 N 为样本总数,初始化概率估计值 $p(x_i)=\frac{1}{2}$.

2 迭代获得强学习器:

FOR $m=1,2,\cdots,M$:

(1)更新响应值及样本权重

$$z_i=\frac{y_i^*-p(x_i)}{p(x_i)(1-p(x_i))}$$

$$w_i=p(x_i)(1-p(x_i))$$

(2)根据权重 w_i 拟合 z_i 关于 x_i 的加权最小二乘回归函数 f_m;

(3)更新强分类器 $F(x)\leftarrow F(x)+\frac{1}{2}f_m(x)$及概率估计

$$p(x)\leftarrow\frac{e^{F(x)}}{e^{F(x)}+e^{-F(x)}};$$

END FOR

3 输出强分类器 $F(x)=sign[\sum_{m=1}^{m}f_m(x)]$

3.2.3　GentleBoost 算法

GentleBoost[7]同样在 1998 年由 Friedman 等提出,GentleBoost 兼具 Real AdaBoost 和 LogitBoost 的特性. 它是在 WTSL 框架下的 Boosting 算法,与 Real AdaBoost 算法流程相同,不同之处在于优化方法选择类似 LogitBoost 的牛顿法. 其算法流程如算法 3.4 所示.

算法 3.4　GentleBoost

输入:训练集 X 及其对应标签 y,迭代次数 M

输出:强分类器 F

1 样本权重初始化 $w_i=1/N, i=1,2,\ldots,N$，其中 N 为样本总数.

2 迭代获得强学习器：

FOR m=1,2,…,M：

(1)根据权重 w_i 拟合 y_i 关于 x_i 的加权最小二乘回归函数 f_m；

(2)更新强分类器 $F(x)\leftarrow F(x)+f_m(x)$；

(3)更新样本权重 $w_i\leftarrow w_i\exp[-y_i f_m(x_i)]$，$i=1,2,\ldots,N$ 并归一化使得 $\sum_i w_i=1$；

END FOR

3 输出强分类器 $F(x)=\mathrm{sign}[\sum_{m=1}^{M} f_m(x)]$

3.2.4 GradientBoost 算法

GradientBoost[8] 由 Friedman 于 1999 年首次提出，它的主要思想是每一次建立的模型，都是之前建立模型的损失函数的梯度下降. 如果模型能够让损失函数持续下降，则说明模型在不断改进，而损失下降最快即损失函数的梯度方向，即每次学习的弱学习器就是梯度的下降方向. GradientBoost 从数值优化的基本点出发，利用最速下降原理，将数值优化泛化到函数空间. 算法流程如算法 3.5 所示.

算法 3.5 GradientBoost

输入：训练集 X 及其对应标签 y，迭代次数 M

输出：强分类器 F

1 定义损失函数 $L(y_i,\rho)$，并将强学习器初始化为

$$F_0(x)=\arg\min_{\rho}\sum_{i=1}^{N}L(y_i,\rho)$$

2 迭代获得强学习器：

FOR m=1,2,…,M：

(1)计算损失函数在当前学习器学习结果处的负梯度值

$$\tilde{y}_i=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)},\ i=1,\cdots,N;$$

(2)更新当前弱学习器

$$a_m=\arg\min_{a,\beta}\sum_{i=1}^{N}[\tilde{y}_i-\beta h(x_i;a_m)]^2$$

(3)更新弱学习器权重

$$\rho_m=\arg\min_{\rho}\sum_{i=1}^{N}L(y_i,F_{m-1}(x_i)+\rho h(x_i;a_m))$$

(4)更新强学习器 $F_m(x)=F_{m-1}(x)+\rho_m h(x;a_m)$；

END FOR

3 输出强分类器 $F_m(x)$.

4 Boosting 算法思想新应用

Boosting 算法自提出以来，出现了很多 Boosting 框架下的变种，如将 Boosting 思想用于多核学习[66]等.

4.1 Boosting 思想用于 kNN

2009 年出现了将 Boosting 思想用于 kNN 的用法，但直接将 Boosting 思想用于 kNN 是不可实现的(由于 kNN 的稳定性，每次学得的弱学习器将会相同或类似)，但 kNN 对输入的选取是敏感的，Nicolás 等[67]提出在输入空间作投影，之后运用 Boosting 框架. 2010 年，Richard 提出比较成型的 Boosting 思想用于 kNN 算法 UNN(universal nearest neighbor)[68]，将最近邻的 k 个样本视作 Boosting 算法中的 k 个弱学习器，并且分配给每个样本相关的权重(即弱学习器权重)，最后加入 Boosting 基本框架迭代学得最终结果. 2014 年，Richard 在 UNN 基础上提出 GNNB[46]，利用 Newton-Raphson 算法最小化特定的凸替代损失. 强学习器的形式为

$$F(x)=\sum_{j?\sim k_x}\alpha_j\times y_j$$

式中，$j\sim k_x$ 表示 j 在 x 的 k 近邻之内，α_j 为弱学习器权重，y_j 为样本标签，即弱学习器. 其算法流程如算法 4.1 所示.

算法 4.1 Gentle Nearest Neighbor Boosting

输入：训练集 $X=\{(x_i,y_i),i=1,2,\cdots,m\}$、类别数目 C 及迭代次数 M

输出：强分类器 F

1 迭代获得强学习器：

FOR c=1,2,…,C：

初始化样本权重 w_i 及弱分类器权重 α_j

FOR m=1,2,…,M：

(1)选择作为弱学习器的样本 j；

(2)计算弱学习器误差 δ_j 并更新样本权重 w_i

(3)更新弱学习器权重 α_j

END FOR

END FOR

2 输出强分类器 $F(x)=\sum_{j?\sim k_x}\alpha_j\times y_j$.

4.2 Boosting 思想用于多元协同表示

协同表示(collaborative representation)通过将数据表示为来自不同子空间样本的集合后再学习[47]. 当数据的一种协同表示不足以或者不能确定样本的最终学习效果时，通过结合数据的多组协同表示可以提高置信度. CRBoosting(collaborative representation boosting)[47]是通过结合多组最优协同表示 CROC(collaborative representation optimized classifier)[69]进行学习的过程. CRBoosting 是在得到学习结果之前分配权重，而 AdaBoost 则是根据学习效果的好坏确定权重. 强学习器形式为

$$H(z)=\arg\min_{i}\sum_{m=1}^{M}\alpha_m\|z-\Psi_i s_i^m\|_2^2$$

式中，α_m 为第 m 个协同表示(弱学习器)的权重，z 为测试样本，s_i^m 为第 i 个子空间的基，Ψ_i 为 z 在 s_i^m 中的协同表示. 其算法流程如算法 4.2 所示.

算法 4.2　CRBoosting

输入：训练集 X、测试集 Y 及其数目 L、对应标签 c_l、变换矩阵 Φ 及迭代次数 M.

输出：协同表示权重 $\{\alpha_m\}_{m=1}^M$

1 初始化样本权重 $D_l(l)=1/L$

2 迭代获得协同表示权重：

FOR m=1,2,…,M:

(1)寻找使得 $|\varepsilon_m/b_m|$ 最大的一组协同表示，其中

$$\varepsilon_m=\mathbb{E}D_m[d_{l,m}],\ b_m=\max_l|d_{l,m}|$$

$$d_{l,m}=\|z_l-\Psi_{c_l}s_{c_l}^t\|_2^2-\min_{i\neq c_l}\|z_l-\Psi_i s_i^t\|_2^2$$

(2)计算当前协同表示的权重

$$\alpha_m=\max\left\{\frac{1}{2b_m}\log\frac{b_m-\varepsilon_m}{b_m+\varepsilon_m},0\right\}$$

(3)更新样本权重

$$D_{m+1}(l)=\frac{D_m(l)}{Z_m}e^{\alpha_m d_{l,m}}$$

END FOR

3. 输出协同表示权重 $\{\alpha_m\}_{m=1}^M$ 并归一化.

4.3　在线 Boosting 算法

2001 年，Oza 等[70-71]提出在线 Boosting 算法[72-74]，并做了大量对比试验，证明在线 Boosting 算法的有效性. 由于离线算法对于存储性能要求严格，2006 年，Grabne 等[75]提出在线 Boosting 算法用于特征选择和特征提取，降低内存消耗的同时，提高算法准确率. 之后的研究包括 Liu 等[76]提出的基于梯度的特征选择在线 Boosting 算法，2008 年 Grabner 等[77]提出半监督在线 Boosting 用于鲁棒追踪等. Chen 等[78]从理论的角度提出在线 Boosting 算法[79]，将用于批处理的弱可学习假设应用到在线中，理论上保证了在弱学习器和样本数量充足时，在线 Boosting 算法可以达到任意精度[80]. Beygelzimer 等[48]将弱可学习的概念发展为弱在线可学习，提出了两种在线 Boosting 算法，是最新的在线 Boosting 算法.

5　结论

Boosting 算法作为当前最为流行的机器学习算法之一，在模式识别、计算机视觉领域有着广泛的应用. 本文从弱可学习的角度出发，回顾了可学习理论与弱可学习理论的基本定义，分析了当前大数据环境下，Boosting 算法面临的数据维数高及数据量大的问题，并指出 Boosting 算法针对大数据环境的解决方案.

Boosting 的优点表现在三个方面，首先，Boosting 不容易出现过拟合，Margin 理论可以很好地解释其合理性；其次，Boosting 参数少，通常情况下，只需确定迭代次数 M；最后，Boosting 算法的学习效果具有逼近贝叶斯最优的性能. Boosting 也存在一些问题，首先，Boosting 不容易出现过拟合的前提是没有 Outlier，Boosting 对 Outlier 是敏感的，这是因为每次的学习效果会影响下次迭代的样本采样分布，使得弱学习器更加注重学习错误的样本；其次，虽然 Boosting 通常情况下只需设定参数 M，但 M 的选取是经验值，在没有 Outlier 的情况下，适当增加 M 的值对学习效果是有效的，但无目的地增大 M 会大大增加计算开销，并且随着 M 的增加，后续的弱学习器效果已经微乎其微；最后，Boosting 算法还没有针对多分类问题的快速有效计算方法，当前主流的用 Boosting 处理多分类问题的三种算法 AdaBoost. M1、AdaBoost. MH 以及二进制编码方法都存在缺陷. 从弱可学习和统计学角度，Boosting 思想的创新可以按照三要素进行，即损失函数选取、权重更新算法以及优化方法；设计高效快速的并行算法，寻求更加紧致的泛化误差界以及收敛速度也是 Boosting 的研究方向之一.

参考文献(References)

[1] KEARNS M J, VALIANT L G. Learning boolean formulae or finite automata is as hard as factoring[R]. Cambridge, USA: Harvard University, 1988.

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

[3] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[4] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning algorithms and an application to boosting [J]. Journal of Popular Culture, 1997, 13(5): 663-671.

[5] FREUND Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121 (2): 256-285.

[6] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[C/OL]//Proceedings of the 13th International Conference on Machine Learning, 1996: 148-156[2015-08-12]. http://citeseerx. ist. psu. edu/viewdoc/download? doi=10. 1. 1. 90. 4143&rep=rep1&type=pdf.

[7] FRIEDMAN J, HASTIE T, TIBSHIRANI R.

Additive logistic regression: a statistical view of boosting[J]. The Annals of Statistics, 2000, 28(2): 374-376.

[8] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. The Annals of Statistics, 2001: 1189-1232.

[9] NAGHIBI T, PFISTER B. A boosting framework on grounds of online learning [C/OL]//Advances in Neural Information Processing Systems, 2014: 2267-2275 [2015-08-12]. http://papers.nips.cc/paper/5512-a-boosting-framework-on-grounds-of-online-learning.pdf.

[10] FREUND Y, IYER R, SCHAPIRE R E, et al. An efficient boosting algorithm for combining preferences [J]. The Journal of Machine Learning Research, 2003, 4: 933-969.

[11] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378.

[12] ESCUDERO G, MÁRQUEZ L, RIGAU G. Boosting applied to word sense disambiguation [C]// Proceedings of the 12th European Conference on Machine Learning. Berlin: Springer, 2000: 129-141.

[13] WEBB G I. Multiboosting: a technique for combining boosting and wagging[J]. Machine Learning, 2000, 40(2): 159-196.

[14] FREUND Y. An adaptive version of the boost by majority algorithm[J]. Machine Learning, 2001, 43(3): 293-318.

[15] BENNETT K P, DEMIRIZ A, MACLIN R. Exploiting unlabeled data in ensemble methods[C]// Proceedings of the Eighth ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 289-296.

[16] DEMIRIZ A, BENNETT K P, SHAWE-TAYLOR J. Linear programming boosting via column generation [J]. Machine Learning, 2002, 46(1/2/3): 225-254.

[17] BÜHLMANN P, YU B. Boosting with the L2-loss: regression and classification [J]. Journal of the American Statistical Association, 2003, 98(462): 324-339.

[18] SERVEDIO R A. Smooth boosting and learning with malicious noise[J]. The Journal of Machine Learning Research, 2003, 4: 633-648.

[19] KÉGL B, WANG L. Boosting on manifolds: adaptive regularization of base classifiers[C/OL]//Advances in Neural Information Processing Systems, 2004: 665-672 [2015-08-12]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.9620&rep=rep1&type=pdf.

[20] HERTZ T, BAR-HILLEL A, WEINSHALL D. Boosting margin based distance functions for clustering [C]//Proceedings of the Twenty-first International Conference on Machine learning. New York: ACM, 2004: 50.

[21] VEZHNEVETS A, VEZHNEVETS V. Modest AdaBoost: teaching AdaBoost to generalize better[J]. Graphicon, 2005, 12(5): 987-997.

[22] HATANO K. Smooth boosting using an information-based criterion [C]//Proceedings of the 17th International Conference on Algorithmic Learning Theory. Berlin: Springer, 2006: 304-318.

[23] WARMUTH M K, LIAO J, R? TSCH G. Totally corrective boosting algorithms that maximize the margin [C]//Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 1001-1008.

[24] BÜHLMANN P, YU B. Sparse boosting[J]. The Journal of Machine Learning Research, 2006, 7: 1001-1024.

[25] BRADLEY J K, SCHAPIRE R E. Filterboost: regression and classification on large datasets [C/OL]//Advances in Neural Information Processing Systems, 2007: 185-192 [2015-08-12]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.404.946&rep=rep1&type=pdf.

[26] RÄTSCH G, WARMUTH M K, GLOCER K A. Boosting algorithms for maximizing the soft margin[C/OL]//Advances in Neural Information Processing Systems, 2007: 1585-1592 [2015-08-12]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.6621&rep=rep1&type=pdf.

[27] MASNADI-SHIRAZI H, VASCONCELOS N. On the design of loss functions for classification: theory, robustness to outliers, and savageboost [C/OL]// Advances in neural information processing systems, 2009: 1049-1056[2015-08-12]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.470&rep=rep1&type=pdf.

[28] FREUND Y. A more robust boosting algorithm[EB/OL]. (2009-05-13)[2015-08-12]. http://aixiv.org/abs/0905.2138.

[29] BÜHLMANN P, HOTHORN T. Twin boosting: improved feature selection and prediction[J]. Statistics and Computing, 2010, 20(2): 119-138.

[30] ZHAI S, XIA T, TAN M, et al. Direct 0-1 loss minimization and margin maximization with boosting [C/OL]//Advances in Neural Information Processing Systems, 2013: 872-880 [2015-08-12]. http://machinelearning.wustl.edu/mlpapers/paper_files/

NIPS2013_5214. pdf.

[31] SHEN C, LIN G, VAN DEN HENGEL A. Structboost: boosting methods for predicting structured output variables[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(10): 2089-2103.

[32] SCHAPIRE R E, SINGER Y. Boostexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2): 135-168.

[33] BERGSTRA J, CASAGRANDE N, ERHAN D, et al. Aggregate features and AdaBoost for music classification[J]. Machine Learning, 2006, 65(2/3): 473-484.

[34] LI F F, FERGUS R, TORRALBA A. Recognizing and learning object categories [A]// Tutorial at International Conference on Computer Vision. Lisbon, Portugal: ACM Press, 2009.

[35] LIU P, HAN S, MENG Z, et al. Facial expression recognition via a boosted deep belief network[C/OL]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 1805-1812[2015-08-12]. http://www. cv-foundation. org/openaccess/content _ cvpr_2014/papers/Liu_Facial_Expression_Recognition _2014_CVPR_paper. pdf.

[36] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: vol. 1. Piscataway: IEEE Press, 2001: 511-518.

[37] SABERIAN M, VASCONCELOS N. Boosting algorithms for detector cascade learning [J]. The Journal of Machine Learning Research, 2014, 15(1): 2569-2605.

[38] FREUND Y, SCHAPIRE R, ABE N. A short introduction to boosting [J]. Journal of Japanese Society For Artificial Intelligence, 1999, 14(50): 771-780.

[39] SCHAPIRE R E. A brief introduction to boosting [C]//Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence: vol. 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 1999: 1401-1406.

[40] SHEN X H, ZHOU Z H, WU J X, et al. Survey of boosting and bagging[J]. Computer Engineering and Application, 2000, 12: 31-32.

[41] SCHAPIRE R E. The boosting approach to machine learning: an overview[M]//Nonlinear estimation and classification. New York: Springer, 2003: 149-171.

[42] LIAO H W, ZHOU D L. Review of AdaBoost and Its Improvement[J]. Computer Systems & Applications, 2012, 21(5): 240-244.

[43] CAO Y, MIAO Q G, LIU J C, et al. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013, 39(6): 745-758.

[44] BÜHLMANN P. Boosting methods: why they can be useful for high-dimensional data[C/OL]//Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC), 2003 [2015-08-12]. http://citeseerx. ist. psu. edu/viewdoc/download? doi =10. 1. 1. 6. 2694&rep=rep1&type=pdf.

[45] SCHAPIRE R E, FREUND Y, BARTLETT P, et al. Boosting the margin: a new explanation for the effectiveness of voting methods [J]. The Annals of Statistics, 1998,26(5): 1651-1686.

[46] NOCK R, ALI W B H, D'AMBROSIO R, et al. Gentle nearest neighbors boosting over proper scoring rules[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 80-93.

[47] CHI Y, PORIKLI F. Classification and boosting with multiple collaborative representations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1519-1531.

[48] BEYGELZIMER A, KALE S, LUO H. Optimal and adaptive algorithms for online boosting [EB/OL]. (2015-02-09) [2015-08-12]. http://arxiv. org/abs/1502. 02651.

[49] VALIANT L G. A theory of the learnable [J]. Communications of the ACM, 1984, 27(11): 1134-1142.

[50] SHALEV-SHWARTZ S, BEN-DAVID S. Understanding machine learning: from theory to algorithms [M]. Cambridge, UK: Cambridge University Press, 2014.

[51] ZHANG T, YU B. Boosting with early stopping: convergence and consistency [J]. The Annals of Statistics, 2005,33(4): 1538-1579.

[52] BAUER E, KOHAVI R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants [J]. Machine Learning, 1999, 36(1): 105-139.

[53] DIETTERICH T G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization [J]. Machine Learning, 2000, 40(2): 139-157.

[54] DUBOUT C, FLEURET F. Adaptive sampling for large scale boosting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1431-1453.

[55] CHI E C, ALLEN G, ZHOU H, et al. Imaging genetics via sparse canonical correlation analysis[C]// 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI). Piscataway: IEEE Press,

2013：740-743.

[56] BREIMAN L. Bias, variance, and arcing classifiers [R/OL]. [2015-08-12]. http://citeseerx.ist.psu.edu/viewdoc/download? doi = 10.1.1.30.8572&rep = rep1&type=pdf.

[57] DRUCKER H, CORTES C. Boosting decision trees [C/OL]//Advances in Neural Information Processing Systems, 1996：479-485 [2015-08-12]. http://papers.nips.cc/paper/1059-boosting-decision-trees.pdf.

[58] QUINLAN J R. Bagging, boosting, and C4.5[C]// Proceedings of the Thirteenth National Conference on Artificial Intelligence. Palo Alto：AAAI Press, 1996：725-730.

[59] BREIMAN L. Prediction games and arcing algorithms [J]. Neural Computation, 1999, 11(7)：1493-1517..

[60] SCHAPIRE R E, FREUND Y, BARTLETT P, et al. Boosting the margin：a new explanation for the effectiveness of voting methods[J]. The Annals of Statistics, 1998,26(5)：1651-1686.

[61] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2)：123-140.

[62] REYZIN L, SCHAPIRE R E. How boosting the margin can also boost classifier complexity [C]// Proceedings of the 23rd International Conference on Machine Learning. New York：ACM, 2006：753-760.

[63] BREIMAN L. Prediction games and arcing classifiers：Technical Report 504 [R]. Berkeley ：University of California, 1997.

[64] ZHOU Z H. Boosting 25 years[R]. Beijing：Institute of Automation, Chinese Academy of Science, 2013.

[65] GAO W, ZHOU Z H. On the doubt about margin explanation of boosting [J]. Artificial Intelligence, 2013, 203：1-18.

[66] TOMER H. Learning distance functions：algorithms and applications[D]. Jerusalem：Hebrew University of Jerusalem, 2006.

[67] GARCÍA-PEDRAJAS N, ORTIZ-BOYER D. Boosting k-nearest neighbor classifier by means of input space projection [J]. Expert Systems with Applications, 2009, 36(7)：10570-10582.

[68] PIRO P, NOCK R, NIELSEN F, et al. Boosting k-NN for categorization of natural scenes [EB/OL]. (2010-01-08) [2015-08-12]. http://arxiv.org/abs/1001.1221.

[69] CHI Y, PORIKLI F. Connecting the dots in multi-class classification：from nearest subspace to collaborative representation [C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway：IEEE Press, 2012：3602-3609.

[70] OZA N C, RUSSELL S. Experimental comparisons of online and batch versions of bagging and boosting [C]//Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York：ACM, 2001：359-364.

[71] OZA N C. Online bagging and boosting[C]// 2005 IEEE International Conference on IEEE Systems, Man and Cybernetics：Vol. 3. Piscataway：IEEE Press, 2005：2340-2345.

[72] WU B, NEVATIA R. Improving part based object detection by unsupervised, online boosting[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2007. Piscataway：IEEE Press, 2007：1-8.

[73] LEISTNER C, SAFFARI A, ROTH P M, et al. On robustness of on-line boosting-a competitive study [C]// 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). Piscataway：IEEE Press, 2009：1362-1369.

[74] BABENKO B, YANG M H, BELONGIE S. A family of online boosting algorithms[C]// 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). Piscataway：IEEE Press, 2009：1346-1353.

[75] GRABNER H, BISCHOF H. On-line boosting and vision[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition：Vol. 1. Piscataway：IEEE Press, 2006：260-267.

[76] LIU X, YU T. Gradient feature selection for online boosting[C] // IEEE 11th International Conference on Computer Vision. Piscataway：IEEE Press, 2007：1-8.

[77] GRABNER H, LEISTNER C, BISCHOF H. Semi-supervised on-line boosting for robust tracking[C]// Proceedings of the 10th European Conference on Computer Vision. Berlin：Springer-Verlag, 2008：234-247.

[78] CHEN S T, LIN H T, LU C J. An online boosting algorithm with theoretical justifications [EB/OL]. (2012-06-27) [2015-08-12]. http://arxiv.org/abs/1206.6422.

[79] LUO H, SCHAPIRE R E. A drifting-games analysis for online learning and applications to boosting [C/OL]//Advances in Neural Information Processing Systems, 2014：1368-1376 [2015-08-12]. http://papers.nips.cc/paper/5469-a-drifting-games-analysis-for-online-learning-and-applications-to-boosting.pdf.

[80] CHEN S T, LIN H T, LU C J. Boosting with online binary learners for the multiclass bandit problem[C/OL]//Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014：342-350 [2015-08-12]. http://machinelearning.wustl.edu/mlpapers/paper_files/icml2014c1_chenb14.pdf.