

Adaboost 算法分类器设计及其应用

许 剑, 张洪伟

(成都信息工程学院, 成都 610225)

摘 要: Adaboost 算法可以将分类效果一般的弱分类器提升为分类效果理想的强分类器, 而且不需要预先知道弱分类器的错误率上限, 这样就可以应用很多分类效果不稳定的算法来作为 Adaboost 算法的弱分类器。由于 BP 神经网络算法自身存在的局限性和对训练样本进行选择的主观性, 其分类精度以及扩展性有待提高。将 Adaboost 算法与 BP 神经网络相结合, 使用神经网络分类模型作为 Adaboost 算法的弱分类器。算法在 matlab 中实现, 对 2 个 UCI 的分类实验数据集进行实验, 结果表明 Adaboost 能有效改善 BP 神经网络的不足, 提高分类正确率和泛化率。

关键词: 弱分类器; 强分类器; BP 神经网络; Adaboost 算法

中图分类号: TP301.6

文献标志码: A

BP 神经网络在分类算法中有广泛的应用, 但是由于 BP 神经网络本质上是梯度下降法, 当遇到比较复杂的目标函数时, 其学习速度和收敛速度会变得很慢, 导致训练失败的可能性较大。而 Adaboost 能够提升任意精度的弱分类器为强分类器, 它的核心思想是通过对一个训练集训练得到多个分类结果, 最后将它们集成起来。在 BP 神经网络中引入 Adaboost 算法作为分类器, 可以弥补 BP 神经网络的缺点, 能提高分类准确率和泛化率。

1 Adaboost 算法概述

1.1 boosting 算法

PAC(probably approximately correct) 是机器学习领域内的早期学习算法。Kearns 和 Valiant 提出, 在 Valiant 的 PAC 模型中, 一个弱学习算法可以被提升为一个具有任意精度的强学习^[1]。令 S 为包含 N 个样本点 $(x_1, y_1), \dots, (x_n, y_n)$ 的样本集, 其中 x_n 是按照某种固定但未知的分布 $D(x)$ 随机独立抽取的。 $Y_n = f(x_n)$, f 属于某个已知的布尔函数集 F 。如果对于任意的 D , 任意的

$0 \ll \varepsilon \ll 1/2, 0 \ll \delta \ll 1/2$, 学习算法能生成一个满足 $P[h(x) \neq f(x)] \leq \varepsilon$ 的概率大于 $1 - \delta$, 并且学习算法的运行时间与 $\frac{1}{\varepsilon}, \frac{1}{\delta}$ 满足多项式关系, 则称为强学习算法^[2]。类似的, 若一个满足 $P[h(x) \neq f(x)] \leq \varepsilon$ 的概率大于 $1/2 + \delta(0 \ll \delta \ll 1/2)$, 只要求这个弱学习算法的正确率大于 50%, 也就是比随机猜测稍好, 则称为弱学习算法^[2]。1990 年, Schapire 使用了构造方法来证明弱学习算法和强学习算法是等价的^[3], 因为强学习算法在通常情况下很难获得, 只要能够找到比随机猜测稍好的弱学习算法, 就可以把它提升为强学习算法。

boosting^[4] 算法由 Schapire 和 Freund 在 1990 年提出, 是最早提出的集成学习算法, 是一种有效提高分类学习系统分类能力的算法。boosting 算法操纵训练样本来产生多个分类假设, 从而建立通过加权投票结合的分类器集合, 其过程如下:

(1) 首先对含有 N 条数据的训练样本进行学习后得到第一个弱分类器;

(2) 将在前面学习过程中分错的样本加上其他新

数据一起构成一个新的含 N 个数据的训练样本集,对这个训练样本集进行训练学习,得到新的分类器;

(3) 把在步骤(1)和(2)中都分错了的样本加上其他新数据构成另一个新的包含 N 个数据的训练样本集,对这个样本集进行训练学习得到新的分类器;

(4) 最后通过对全部得到的弱分类器进行加权投票表决得到强分类器,即某样本被分为哪一类要通过全部弱分类器的加权投票表决。

boosting 算法需要解决两个主要问题:

(1) 怎样调整训练集里的样本分布,从而使训练集产生适合的弱分类器;

(2) 怎么把训练所得弱分类器集成得到效果更好的强分类器。

1.2 Adaboost 算法

Adaboost 算法是 Schapire 和 Freund 在 1995 年提出的,其关键思想是针对同一个训练集训练多个的弱分类器,最后将这些弱分类器集成起来构成最终的强分类器^[5]。Adaboost 算法主要是根据每次训练过程中样本集内每个样本的分类结果是否正确来改变数据样本分布,即是修改样本的权值。修改过权值的新数据集再次进行训练得到新的弱分类器,最后通过某种策略,常用的如投票加权方式,将得到的弱分类器集成起来生成最后的决策分类器。Adaboost 分类算法可以过滤掉非关键的数据特征,以减小数据维度。弱学习过程得到的弱假设影响最后的强假设的正确率,它有效地解决了早期 boost 算法在实际运用中的困难,因此更能全面地挖掘学习算法的能力,因此叫 adaptive boosting,简称 Adaboost。

Adaboost 算法是 Freund 和 Schapire 根据已经存在的在线分配算法提出的。跟 boosting 算法最大的不同的是:Adaboost 算法不需要预先知道弱学习算法的误差精度,并且最后得到的强分类器的分类精度依赖于所有弱分类器的分类精度影响得到的强分类器的分类精度,这样就可以去寻找要求更低的分类算法。

Adaboost 算法初始状态下每个样本的权重是相同的,对此样本分布训练得到一个弱分类器。对于被分错的样本,增加权重;相反情况下,降低权重。这样,被分错的样本就更能出现在新的样本分布中,这些分错的样本被着重训练。在经过 N 次训练后,我们得到了 N 个弱分类器,最后把这 N 个弱分类器按一定的权重集成起来得到最终想要的强分类器^[6]。

Adaboost 算法不断加入新的弱分类器以达到足够小的误差率。在 Adaboost 算法中,每个训练样本都被赋予一个权重系数,来表明它被下一次选入一个训练过程的概率。当某个样本点在当前训练过程中被正确地分类,那么应该降低它的权重以降低选择该样本进入下次训练过程的几率;相反,如果某个样本在当前训练工程中被错误的分类,那么应该增加它的权重以使得它在下个弱分类器构造时更能被选中。这样 Adaboost 就能更关注那些分类困难的样本,提高最后算法分类结果的正确率^[7]。

2 adaboost_bp 神经网络分类算法

2.1 算法思路

Adaboost 算法是一种迭代算法。目前,对 Adaboost 算法的研究以及应用大多集中于分类问题,同时近年也出现了一些在回归问题上的应用^[8]。Adaboost 算法能够提高任意给定弱分类器的分类精度,因此,本文针对 BP 神经网络自身的局限性和训练样本选择的主观因素,为提高其分类精度,将 Adaboost 算法与 BP 神经网络相结合,建立了 adaboost_bp 神经网络分类模型^[9]。模型采用 BP 神经网络作为弱分类器,根据每次训练样本分类的优劣,减少或增加其对应的权重,并使用改变权重后的样本重新对弱分类器进行训练,最后将这些弱分类器的训练结果进行集成,得到最终的输出。

2.2 算法实现步骤

算法基本步骤如下^[10]:

步骤 1 数据选择和网络初始化。

从样本空间选择 m 组训练样本,初始化训练样本的分布权值 $D_1(i) = \frac{1}{m}, i = 1, 2, \dots, m$, 根据样本输入输出维数确定神经网络结构,初始 bp 神经网络权值和阈值。

步骤 2 弱分类器分类。

训练第 t 个弱分类器时,用训练样本 $g(t)$ 训练 BP 神经网络,并且分类训练样本输出,得到训练样本的分类误差和 e_t 。

$$e_t = \sum D_t(i) [h_t(x_i) \neq y_i] \quad (1)$$

步骤 3 计算分类序列权重。

根据训练样本 $g(t)$ 的分类误差和 e_t 计算权重 a_t :

$$a_i = \frac{1}{2 \ln \left(\frac{1 - e_i}{e_i} \right)} \quad (2)$$

步骤4 测试数据权重调整。

根据分类序列权重 a_i 调整下一轮训练样本的权重,其数学表达式为:

$$D_{i+1}(i) = \frac{D_i(i)}{B_i} \times \exp[-a_i y_i h_i(x_i)] \quad (3)$$

步骤5 强分类函数。

将得到的 T 个弱分类器的权重 a_i 归一化,则强分类函数分类结果 $H(x)$ 为:

$$H(x) = \text{sign} \left(\sum_{i=1}^T a_i h_i(x) \right) \quad (4)$$

式中, $h(x)$ 为 T 个弱分类器得到的分类样本的分类值。

3 实验与结果分析

3.1 实验数据

为了验证本文提出算法的有效性和实用性,将用 UCI Machine Learning Repository 中的 Iris Dataset 和 Breast Cancer Wisconsin Dataset 数据集来进行实验,并对数据都采用归一化处理,数据归一化把数据都转化为 $[0,1]$ 之间的数,以便消除各维数据间的数量级差别,避免数据各维度之间由于数量级差别过大而造成误差过大。常用的归一化方法为:

$$x_i = \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})}$$

式中 x_{\max} 为数据集每一维的最大值, x_{\min} 为数据集每一维的最小值。

3.2 实验结果与分析

3.2.1 Iris Dataset

鸢尾花数据是模式识别文献中最著名的数据集(表1),该数据集包含3个类别的鸢尾花数据,每个类别50条数据,总共有150条数据。一个类与其他两个类是线性可分的,而后两个不是线性可分的,所以通常用来检测分类器的效果。该数据集包含4个数值属性和1个类别标签,实验结果见表2。

表1 鸢尾花实验数据描述

项目	最小值	最大值	均值	标准差	类相关
萼片长度	4.3	7.9	5.84	0.83	0.7826
萼片宽度	2	4.4	3.05	0.43	-0.4194
花瓣长度	1	6.9	3.76	1.76	0.949
花瓣宽度	0.1	2.5	1.2	0.76	0.9565

表2 鸢尾花实验结果对比

算法	正确率
BP 算法	94.5%
Adaboost	97.2%

3.2.2 Breast Cancer Wisconsin Dataset

威斯康辛乳腺癌诊断数据是另外一个著名的数据集(表3)。该数据集包含2个类别(良性、恶性)的乳腺癌诊断数据,总共有699条数据,其中良性458条,恶性241条。该数据集包含10个数值属性(其中第一个数据为样本编码号在数据集使用中省略)和1个类别标签。实验结果见表4。

表3 乳腺癌诊断实验数据描述

序号	属性	域
1	样本编码号	ID号
2	肿块厚度	1-10
3	细胞大小的均匀性	1-10
4	细胞形状的一致性	1-10
5	边缘粘连	1-10
6	单个上皮细胞大小	1-10
7	裸细胞核	1-10
8	湿性染色质	1-10
9	正常核仁	1-10
10	有丝分裂	1-10
11	类	(2为良性,4为恶性)

表4 乳腺癌诊断实验结果对比

算法	平均正确率
BP 算法	95.6%
Adaboost	98.3%

4 结束语

针对BP神经网络训练时间较长以及可能训练失败,提出把BP神经网络引入Adaboost算法改进分类算法。该算法将BP神经网络作为弱分类器,经过反复训练的弱分类器组合起来称为一个强分类器。实验结果表明,该算法比BP神经网络能有更低的分类误差以及较好的扩展性。

参考文献:

- [1] Valiant L G. A theory of learnable[J]. Communications of the ACM, 1984, 27(11): 1134-1142.
- [2] Kearns M J, Valiant L G. Cryptographic limitations on

- learning Boolean formulae and finite automata[J]. Journal of the ACM(JACM),1994,41(1):67-95.
- [3] Freund Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121 (2): 256-285.
- [4] Schapire R E. A brief introduction to boosting[C]// Proceedings of the 16th international joint conference on Artificial intelligence, Stockholm, Sweden, July 31-August 6, 1999: 1401-1406.
- [5] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]// Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, July 3-6, 1996: 148-156.
- [6] 涂承胜, 刁力力, 鲁明羽, 等. Boosting 家族 AdaBoost 系列代表算法[J]. 计算机科学, 2003, 30(3): 30-34, 145.
- [7] 付忠良. 关于 AdaBoost 有效性的分析[J]. 计算机研究与发展, 2008, 45(10): 1747-1755.
- [8] 曹莹, 苗启广, 刘家辰, 等. AdaBoost 算法研究进展与展望[J]. 自动化学报, 2013, 39(6): 745-758.
- [9] 董元元, 陈基漓, 唐小侠. 基于 BP_Adaboost 的文本分类研究[J]. 网络安全技术与应用, 2012(3): 42-43.
- [10] 李睿, 张九蕊, 毛莉. 基于 AdaBoost 的弱分类器选择和整合算法[J]. 兰州理工大学学报, 2012, 38(2): 87-90.

Design and Application of Adaboost Algorithm Classifier

XU Jian, ZHANG Hongwei

(Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Adaboost algorithm can promote a weak classifier to a strong classifier without knowing the error rate upper limit of the weak classifier in advance, so a lot of classifiers which are not so stable can be used as weak classifiers in Adaboost algorithm. Because of the limitation and subjectivity in training samples selection of the BP neural network algorithm, its classification accuracy and scalability need to be improved. So the Adaboost algorithm is combined with BP neural network, in which the neural network classification model is used as a weak classifier. Algorithm is realized in matlab, and two UCI data sets is used to do the experiment. The results show that Adaboost can effectively overcome the shortcomings of BP neural network, improve the classification accuracy and the rate of generalization.

Key words: weak classifier; strong classifier; BP Neural Network; Adaboost algorithm