北京师范大学珠海分校

本科生毕业论文

论文题目　Boosting集成学习算法研究

学　　院　应用数学学院

专　　业　数学与应用数学

学　　号　1717010106

学 生 姓 名　章俊鑫

指导教师姓名　李艳

指导教师单位　北师大珠海分校应用数学学院

2020年 11月25日

Boosting 集成学习算法研究

文献综述

1.1 Boosting 集成学习算法发展历史

在人工智能这一领域，自 1956 年提出人工智能这一概念发展至今，机器学习在人工智能这一学科中一直都是发展最快的分支之一，同时也是在人工智能中最能够凸显机器能够具有独立自主学习的能力以及分析思考的智能。[1]伴随着软件、硬件以及需求的发展，尤其在现如今大数据时代的来临，机器学习这一学科的存在对于这一时代的意义显得尤为重要。而在机器学习领域中，自发展至今，如何提高模型训练的精度一直都是人们最为关心的部分。人工智能中机器学习的研究者们一直都在致力于如何提高通过已知样本已经训练好的模型对新的测试样本得到一个尽可能高的精度估计，然而经过大量的算法研究者的努力以及付出。至今还是认为并没有哪一种算法能够非常好的构造一种模型使得测试精度大大提升，虽然优化后的模型算法精度确实在一定程度上能够提高，但是能够提高的上限非常局限，想要通过目前已有的算法构建出一个高精度模型估计依旧是非常困难的事情。但是换个角度去思考，以现在的算法构建出多个精度略低的模型却是绰绰有余，如果能够集成多个精度略低的模型来得到一个集成后能够提升精度的模型，提高模型精度的想法在通过集成这一层面来实现，Boosting 算法应运而生。[2]

Boosting 算法是一种通过诸如集成模型和加权系数等算法提高给定学习算法准确性的方法。这一说法源自于 Valiant 在二十世纪初所提到的的 PAC 学习模型。[3]在当时 Valiant 首次提出了弱机器学习和强机器学习的基本概念，他在书中第一次给出了定义在机器学习中识别中准确率高于 50%。也就是说，如果存在一种机器学习算法并且它的准确率能够稍微高于人类行为层面上的随意猜测的统计，比如机器测算抛骰子落地后顶端是哪一面的准确率能够稍高于由人类随意猜测后统计的概论那么在人工智能这一领域中机器学习这一学科中统一认为这种类型的算法是弱机器学习算法反之称之为强机器学习算法。而与此同时，Valiant 在这一基础上提出了一个更为大胆的想法，即弱机器学习算法能否通过 PAC 学习模型中的算法实现弱学习算法的模型精度提升。在弱机器随机学习算法中，有些算法的模型准确度要比弱随机机器学习算法稍强一些。如果能够实现，那么就不必费尽心思去构造强机器学习算法，只需要通过集成模型来改进弱机器学习算法的模型精度即可达到最终目的。

1990 年，Schapire 最先构造出一种多项式级的算法，即最初的 Boosting 算法。这种算法可以将弱分类规则转化为强分类规则。[4]

1993 年，Drucker 和 Schapire 第一次以神经网络作为弱学习器，应用 Boosting 算法来解决实际的 ORC 问题。由于早期的 Boosting 算法在解决实际问

题时要求事先知道弱学习算法学习正确率的下限，这一步本身就很难实现。[5]到了二十世纪尾叶，首次提出并且构造了一个多项式级的算法，并且该问题已经被机器学习的研究者们进行了充分的实践证明。这是最早的正真意义上的Boosting集成学习算法。而后又过了一段时间，德国的Freund对原始的Boosting集成学习算法进行了优化提出了一种更有效的boosting集成学习算法。二十世纪末，Boosting算法再一次进行了算法上额优化以及改进，尤其是提出了AdaBoost（自适应boosting）这一划时代的加权算法。该集成学习算法的效能相比最原始的Boosting算法提高了一些，但最具意义的是该集成学习算法不需要弱机器学习具备任何先验知识，这使得Boosting算法的受众性更强，对于实际的问题更容易应用上去。

之后，随着投票权重这一想法对于Boosting这一集成学习的算法进行优化，推出了AdaBoost M1和AdaBoost M2等算法，使得Boosting集成学习算法受众更广，更实用，并且精度也随之上升，到目前为止，boosting集成学习算法在机器学习领域备受关注。[6]

1.2 Boosting集成学习算法研究意义

随着时代的发展，机器学习这一领域在信息，医疗，通信，各行各行都有着急速的发展以及使用，各个学科领域对于机器学习算法的需求日益增加，本篇论文通过对已有的Boosting集成学习算法的文献再研究希望从中整合或并找出可能被忽略的但是能够在一定程度上提高集成后模型精度的部分，也可在算法上实现一定程度的优化，也可以通过叠加算法通过迭代使得模型精度上升，又或者在原有基础上实现一定程度的算法整合或是创新以实现最终集成模型精度的再提升，使得Boosting的集成学习算法在机器学习这一学科的应用上更上一层楼，用途更加广泛，让数据能够通过不同的聚类、分类、神经网络等等，不同机器学习算法的剖析方式来使得各行各业目前在模型精度的瓶颈能够有一个细微的提升，为社会为国家在信息技术，人工智能等领域做出更多的贡献，为人民的信息生活提供更加优质的服务以及技术。

1.3 Boosting集成学习算法在国内外研究现状

1.3.1 国内研究现状

2012年高敬阳对在实践中最典型最有应用价值的 Adaboost 算法进行了一定程度的修正。AdaBoost 算法有错误样本恶性积累的缺点。随着迭代的继续，错误样本的权重呈指数级不断上升，便会出现恶性积累，这种恶性积累将会一直持续下去。高敬阳为了避免这种恶性积累的产生，以及造成的过拟合现象对算法进行了修改，针对 AdaBoost 算法权值修改策略中存在过分偏重于困难样本的情况他提出了基于争议度修改权值的算法 ERstd—AdaBoost；针对 AdaBoost 算法产生的过拟合现象，提出了基于样本分布调整权值的算法 ABSD；针对逆向权值

分布策略的集成网络泛化性能、个体分类器泛化性能及网络的差异度进行了深入研究，提出了逆向权值分布策略的改进算法 IB+。[7]

2014 年卢婷研究组合分类器中的算法。在该算法中，每个样本被赋予一个权重，这个权重代表该样本被选入训练子集的概率。迭代过程中，如果一个样本前一次被正确分类，那么它的权重就会增大，反之权重减小。通过这种方式，算法聚焦于那些难分的样本，从而提高困难样本的分类正确率。她使用算法对不平衡数据集进行分类。用、伪逆、和这种分类器作为的基分类器，实验对比分析了这种组合分类器对少数类分类正确率的影响，以及对所有样本性能的影响，得到了一些有益的结论。本文对算法进行了改进，不再固定训练子集的大小，而是根据每个样本的权重和训练样本集容量的乘积上取整结果，决定每个样本被选入新训练子集的次数。一方面，使得训练子集都包含了所有的样本，没有信息遗失，提高了分类性能。另一方面，避免了训练子集中某一类别样本数目很多，其它类别样本数目很少甚至没有的情况从而有效避免了过拟合和偏见问题。[8]

2016 年蔡小龙认为在 Boosting 算法中由于传统的学习算法多是基于 ERM(经验风险最小化）的原则，而 ERM 原则是不适定的，会发生过拟合现象即用过于复杂的函数去巧合有限样本的情形，推广性较差。为了避免此类问题的发生，蔡小龙研巧的是正则化框架下的 Boosting 算法。针对独立同分布样本的情形，证明了 Boosting 算法是一致的。因为无论从理论上，还是实际应用中，独立同分布的条件或假设都是很强的，故我们又针对非独立同分布，即混合序列样本的情形，证明了基于混合序列的正则化 Boosting 算法是一致的。[9]

2019 年王超提出一种基于分位逻辑回归思想的逻辑树模型。该模型针对加权样本分类问题能够更好地利用样本权重信息。然后，对现有的 Boosting 算法进行分析，提出一种更加平缓的损失函数，并基于此提出两种新的 Boosting 算法。新的 Boosting 算法能够较好地避免过拟合问题，并且由于对错误样本权重分配力度呈线性增长，与指数损失相比，该模型具有较好的稳健性。模拟数据和实际数据均表明，本文提出的逻辑树模型能够更好地适应协变量个数为 10-50 的分类问题；基于 Ada.Boost.MH 思想将两种 Boosting 算法推广到多分类问题。[10]

在集成分类中，如何对基分类器实现动态更新和为基分类器分配合适的权值一直是研究的重点。随后就是今年 2020 年 11 月份杜诗语等人针对以上两点，提出了 BIE 算法和 BIWE 算法。BIE 算法通过最新训练的基分类器的准确率确定集成是否需要替换性能较差的基分类器及需替换的个数，实现对集成分类器的动态迭代更新。BIWE 算法在此基础上提出了一个加权函数，对具有不同参数特征的数据流可以有针对性的获得基分类器的最佳权值，从而提升集成分类器的整体性能。实验结果表明，BIE 算法相较对比算法在准确率持平或略高的情况下，可以减少生成树的叶子数、节点数和树的深度；BIWE 算法相较对比算法不仅准确率较高，而且能大幅度减少生成树的规模。[11]

参考文献

[1] 周志华. 机器学习[M].北京：清华大学出版社,2016.

[2] 于玲,吴铁军. 集成学习:Boosting 算法综述[J].模式识别与人工智能,2004,17(01):52-59.

[3] Valiant L.G. A Theory of the learnable[P].Communications of the ACM, 1984,27(11):1134-1142.

[4] Schapire R.E. The Strength of Weak Learnability[J].Machine Learning,1990,5(2):197-227

[5] Drucker H,Schapire R.E,Simard P. Boosting Performance in Neural Networks[J]. International Journal of Pattern Recognition and Artificial Intelligence,1993,7(4):705-719

[6] 董乐红,耿国华,高原. Boosting 算法综述[J].计算机应用与软件,2006(08):27-29.

[7] 高敬阳. 神经网络集成 BOOSTING 类算法研究[D].北京化工大学,2012.

[8] 卢婷. 基于 AdaBoost 的分类器学习算法比较研究[D].华东理工大学,2014.

[9] 蔡小龙. 正则化 Boosting 算法的一致性[D].湖北大学,2016.

[10] 王超. 加权样本分类算法设计和基于加法逻辑回归模型的 Boosting 算法设计[D].华中师范大学,2019.

[11] 杜诗语,韩萌,申明尧,张春砚,孙蕊.基于 Boosting 的迭代加权集成分类算法[J/OL].计算机应用研究:1-7[2020-12-13].https://doi.org/10.19734/j.issn.1001-3695.2020.04.0101.