北京师范大学珠海分校本科生毕业论文（设计）开题报告

学院（部）：

|  |  |  |  |
| --- | --- | --- | --- |
| 毕业论文(设计)题目 | 分类算法的集成学习 | | |
| 姓名 | 吕梦恬 | 学号 | 1717010057 |
| 专业 | 数学与应用数学 | 指导教师 | 李艳 |
| 开题报告内容 | | | |
| 一、选题背景及选题意义、国内外研究现状、初步设想及拟解决的问题：  **选题背景及意义：**  20世纪80年代，大数据概念首次提出。2011年以麦肯锡咨询公司发布《大数据：下一个创新、竞争和生产率的前沿》报告为标志，大数据得到广泛的研究和应用。在大数据研究热潮中，数据挖掘技术是大数据分析的核心和骨干技术之一，也是大数据时代最重要的研究领域之一。作为一个多学科交叉研究领域，数据挖掘融合了数据库技术、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等多种技术的研究成果。数据挖掘是指在大型数据存储库中，自动或半自动地发现有用信息的过程从大型数据中提取隐藏的、前所未知的、潜在有用的信息或模式，数据挖掘任务主要是预测和描述，往下细分可大致分为分类、回归、聚类和关联分析这四种。由此可见，在数据挖掘技术中，分类任务占据着十分重要的研究地位。[1]  分类任务是指给定每条数据包含一组属性且其中一个属性为类标的训练集，利用这个训练集训练一个模型，使得这个模型能将未知数据尽可能准确分配到一个类。而分类算法是指由一群已知类标的样例训练出一个分类学习器的算法，使得该分类算法可以用训练出的分类器对未来的或未知的样本和数据进行分类预测。常见的分类算法有KNN算法、朴素贝叶斯算法、决策树算法、逻辑回归和SVM等。现如今，分类广泛应用于医疗诊断[2][3]、信用卡系统的信用分级、图像模式识别[4][5]等方面。[1]  由于计算机技术的不断发展，需要处理的数据量已经远大于从前，某些单一的分类学习器的泛化性能有所下降，为了能得到泛化性能更强的学习器，集成学习方法应运而生。集成学习通过按照设定的策略构建并结合多个相同或不同的学习器来完成学习任务，即先产生一组由一个现有的学习算法从训练数据产生的个体学习器，再用某种策略将它们结合起来，输出一个比单一学习器泛化性能要强的学习器。根据个体学习器的不同，集成方法主要分为以Boosting为代表的个体学习器间存在强依赖关系、必须串行生成的序列化方法和以Bagging为代表的个体学习器间不存在强依赖关系、可同时生成的并行化方法。根据以往学习的一般经验，集成学习把多个学习器结合起来，最终将获得比最好的单一学习器更好的性能，即一般情况下集成后得到的精确度会更高。[6]  目前，集成学习更重视训练模型的精度与多样性。本文通过对比集成前后的分类学习器的精度，以及对比不同方法集成得到的分类器的精度，分析得到何种分类算法集成后精度更高以及何种方法集成精度更高。  大数据时代的到来，为世界带来了大数据科学技术，机器学习和数据挖掘也紧随发展的脚步不断发展。针对分类问题，研究基于集成学习的分类算法，主要是研究何种分类算法可以通过何种方法的集成学习得到精度更高的算法，根据此研究，为人们以后研究不同领域应用或创新分类算法时提供一个更高效更有力的方法和规律，尽可能得到重大突破，大大降低了未来研究过程中因精度过低而导致的整体研究性能低下的可能性，以便于大数据科学技术与应用发展快速前进，有助于人类研究紧跟时代脚步。  **国内外研究现状：**  2019年，袁泉、郭江帆和赵学华[7]提出了一种基于集成学习的不均衡数据流的分类算法，为了处理目标变量在随着时间的推移的过程中发生一些不可预见的现象导致模型的预测精度变低，即概念漂移这一问题，设计出了ECIDS算法。这个算法将数据流切分成一定数量的数据块，然后对应训练出相应数量的基分类器，计算基分类器的性能度量函数得出该基分类器的权值并添加到集成分类器中，当集成分类器中的基分类器数量达到预设值时，将后面添加的性能好的基分类器替换集成分类器中性能差的基分类器，最终采用加权投票策略进行预测分析，以此来应对概念漂移和不均衡数据流问题。  2020年8月，陈伟宏[8]提出了一种新的混合降维集成学习方法HDRF来提高集成系统中各分支之间的多样性，同时保留了各个分支更多的训练样本信息。  2020年11月，杜诗语、韩萌等人[9]针对如何实现集成分类中基分类器的动态更新以及如何分配恰当的权值提出了基于Boosting的BIE算法和BIWE算法。前者通过训练后得到的基分类器的精度确定是否替换性能较差的基分类器，从而实现动态更新问题。后者在前者的基础上添加了一个加权函数，计算获得基分类器最终权值，进而提升集成分类后得到的模型的性能。同年，Adel S. Assiri、Saima Nazir和Sergio A. Velastin[10]提出了一种乳腺肿瘤综合分类方法，他们评估了八种分类算法的性能以及投票法和平均法的策略性能，根据F3分数将性能最佳的三种分类算法（简单逻辑回归学习、随机梯度下降优化SVM学习和多层感知器网络）的预测结果用于采用了不加权多数投票策略的集成学习，提出在未来将计划评估不同特征选择算法，确定最小特征子集，帮助准确将乳腺癌分为良性或恶性。  2021年初，朱亮、徐华和崔鑫[11]发现对基分类器在系数和多样性上进行改进有利于提升传统AdaBoost算法性能，因此引入与AdaBoost相关性最高的双误度量DF，优化基分类器系数的计算，提出了WD AdaBoost算法。  参考文献：   1. 毛国君,段立娟.数据挖掘原理与算法[M].北京.清华大学出版社.1,165.2016 2. 汪琳琳,沈璐,施俊,费晓燕,周玮珺,徐浩煜,刘立庄.基于自步学习的多经验核映射集成分类器在乳腺癌超声计算机辅助诊断上的应用[J/OL].生物医学工程学杂志:1-9.2021 3. 邓卓,苏秉华,张凯.基于集成学习的乳腺癌分类研究[J].中国医疗设备,35(12):59-62.2020 4. 宝音图,刘伟,牛朝阳,李润生,张浩波.联合集成学习与EfficientNet的光学遥感图像场景分类[J/OL].计算机工程: 1-11.2021 5. 吴建,许镜,丁韬.基于集成迁移学习的细粒度图像分类算法[J].重庆邮电大学学报(自然科学版),32(03):452-458.2020 6. 周志华.机器学习[M].北京.清华大学出版社.53,196.2016 7. 袁泉,郭江帆,赵学华.一种基于集成的不均衡数据流分类算法[J].计算机工程与科学,41(08):1519-1524.2019 8. 陈伟宏.基于分类器选择的集成学习算法研究[D].华南理工大学.2020. 9. 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.2020 10. Adel S. Assiri, Saima Nazir, Sergio A. Velastin. Breast Tumor Classification Using an Ensemble Machine Learning Method. Journal of Imaging: 6(6). 2020 11. 朱亮,徐华,崔鑫.基于分类器系数和多样性的改进AdaBoost算法[J/OL].计算机应用:1-11.2021   **初步设想及拟解决的问题：**   1. 首先简单介绍分类算法和集成学习的相关理论和代表算法； 2. 对主要几种分类算法进行评估，取预测精度最好的三个或五个算法作为基分类器； 3. 针对主要的两种集成方法Bagging和Boosting分别实验，得到预测精度最佳的集成分类器。 | | | |
| 1. 论文撰写过程中拟采取的方法和手段： 2. 通过查阅相关书籍和资料，获取相关的理论知识； 3. 自行编写本文需要的代码进行实验，得到实验结果和所需图表； 4. 根据实验结果进行对比，得出所需结论。 | | | |
| 三、论文撰写提纲：  题目  作者  摘要  关键词  1 绪论  1.1 研究背景  1.2 选题意义  1.3 国内外研究进度  2 理论概念  2.1分类算法  2.2集成学习  3 数据来源  4 实验  4.1实验结果  4.2对比分析  5 结论  6 参考文献  7 附录：主要是程序代码，以及一些公式验算  致谢 | | | |
| 1. 计划进度：   1月：提交开题报告  2月：写初稿  3月初：提交初稿  3月初-3月底：修改论文，完成最终稿 | | | |
| 五、指导教师意见：  指导教师签名：  　　　　　 年 月 日 | | | |

|  |
| --- |
| 六、教学院（部）长意见：  教学院（部）长签名：  （加盖学院或学部章） 　 年 月 日 |

注：本表一至四由学生填写；五、六由指导教师和学院填写