**华东师范大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **论文题目** | **基于机器学习的 OpenRank 指标拟合与优化** | | | | | | 数据科学与工程学院 | | | | 数据科学与大数据技术专业 | | | | | | | |
| 学生姓名 | | 方蕴仪 | | 学号 | | | 10205501437 | | | | |
| **一、选题的背景与意义** | | | | | | | | | | | | | | | | | | |
| 如今，开源作为一种将软件的源代码向公众开放的开发模式，以一种允许任何人查看、使用、修改和分享代码的方式促进了计算机从业者之间的合作与共享，为计算机科学和代码工作者社区带来了重大的影响，也为计算机的社区合作、知识传播和技术发展提供了有力的支持。而项目全域 OpenRank 指标是在整个 GitHub 上的开发者与仓库的协作关系网络之上构建的、基于 OpenRank 算法得到的一个具有时序信息的网络指标，它可以有效地体现出一个项目在开源生态中的协作影响力。  基于机器学习的对 OpenRank 指标进行拟合与优化具有重要的意义，这主要体现在以下几个方面：  首先，OpenRank 指标是基于整个 GitHub 上的协作关系网络和时序信息生成得出的，这使得它具有规模巨大、结构复杂的特点。机器学习算法能够通过对这些庞大而多样化的数据进行分析和学习，捕捉到其中潜在的模式和规律。通过拟合，我们能够更加准确地理解 OpenRank 指标与不同项目、开发者以及仓库之间的关系，进而提高对项目在开源生态中的协作影响力的评估准确度。  其次，机器学习的优化过程可以帮助调整模型参数，从而更好地适应不断变化的开源生态。由于开源社区的动态性，项目和开发者之间的关系随时可能发生变化。通过机器学习的实时拟合与优化，我们能够对这些变化做出迅速、直接的反应，确保 OpenRank 指标的时效性和灵活性。  除此之外，机器学习的应用还能够为 OpenRank 指标的未来预测提供有力支持。通过对历史数据的学习，机器学习模型可以识别趋势和模式，为项目在未来的发展趋势提供预测性的分析。这有助于开发者和项目管理者更好地制定策略和决策，以更好地适应开源生态的演变。  总体而言，基于机器学习的拟合与优化不仅可以提高 OpenRank 指标的准确性和实时性，还能够为开源项目的未来规划提供智能支持，促进整个开源社区的可持续发展。这一方法使得我们能够更全面、深入地理解开源协作网络的复杂动态，从而为开发者与项目提供更好的支持。 | | | | | | | | | | | | | | | | | | |
| **二、研究的主要内容和预期目标** | | | | | | | | | | | | | | | | | | |
| 1. **指标的拟合与优化**   在这一部分，我计划以拟合精度为检验目标，使用 CHAOSS 统计型数据指标来拟合 OpenRank 网络指标。  在数据收集方面，我计划选用 OpenDigger 项目中提供的时间序列数据与 GitHub 仓库中得到的异质图数据：包括仓库活跃度数据、 开发者活动数据、与代码贡献数据在内的时间序列数据可以用于描述仓库的历史状态，开发者的活动状况，代码贡献等各方面的信息；而 GitHub仓库中得到的issue，PR和actor等信息得到的异质图数据则能够反应出开发者、issue和PR之间的交互关系。  在算法方面，我计划尝试多种机器学习模型与超参数的组合，并从中选取效果较优的组合作为得到最终结果的模型，我计划尝试经典的机器学习模型包括线性回归模型、决策树模型等，以及在近年被提出的性能效果均较为优异的机器学习算法XGBoost与LightGBM等。   1. **结果的可视化**   开源领域的可视化大屏是一种基于开源技术实现的数据可视化解决方案，主要用于展示大规模数据的实时状态、趋势和关系，以便用户能够快速准确地了解数据的整体状况，从而做出更好的决策。我计划实现一个动态展示项目指标数据与对指标进行预测得到结果的可视化大屏，以直观、清晰的方式对项目结果进行一个展现。 | | | | | | | | | | | | | | | | | | |
| **三、拟采用的研究方法、步骤** | | | | | | | | | | | | | | | | | | |
| 1. **数据预处理**   我们将json格式的时间序列数据转化为csv格式，将异质图数据转换为数值型数据，并对完成格式转换的数据进行统一格式、补充缺失值等方式的数据清洗。此后对数据进行特征融合，通过结合现有特征，创造新的特征以增强模型的预测能力。   1. **数据分析**   对数据进行相关性分析与聚类分析，并且对重新整合得到的代表各metrics的csv文件进行聚类处理，通过利用过去的特征值来影响当前的特征值，从而更好地捕捉时间相关性和趋势。   1. **数据的拟合与优化**   将归一化后的数据放入不同的机器学习模型与超参数组合中进行拟合与超参数调优， 在模型训练完成后，我们需要通过评估指标量化模型的预测能力、评估模型的性能，并得到表现最佳的模型与超参数组合。   1. **结果可视化**   我计划使用Vue3与Echarts作为前端的主要工具，实现高交互性、高可用性的智能可视化大屏，全面展示开源项目的情况。 | | | | | | | | | | | | | | | | | | |
| **四、研究的总体安排与进度** | | | | | | | | | | | | | | | | | | |
| 2023年12月底前：完成项目开题与前期准备工作  2024年1月至2024年3月初：完成指标拟合与调优  2024年3月至2024年4月初：完成可视化大屏的制作  2024年4月至2024年5月：完成毕业论文的写作 | | | | | | | | | | | | | | | | | | |
| **五、参考文献** | | | | | | | | | | | | | | | | | | |
| 1. **XGBoost: A Scalable Tree Boosting System(**[1603.02754.pdf (arxiv.org)](https://arxiv.org/pdf/1603.02754.pdf)) 2. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree(**[LightGBM: A Highly Efficient Gradient Boosting Decision Tree (neurips.cc)](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)) | | | | | | | | | | | | | | | | | | | |
| **论文题目** | |  | | | | | | 学院 | | | | | 专业 | | | | | |
| 学生姓名 | |  | | | 学号 | |  | | | |
| **六、指导教师意见** | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | |
|  | | | | | 签字： |  | | | | 年 | | 月 | | | | 日 |  | |
| **七、开题答辩小组意见** | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | |
|  | | | 小组成员签字： |  | | | | | | 年 | | 月 | | | | 日 | |  |

教务处编制