北 京 交 通 大 学

课程设计开题报告表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 课程名称 | | 大数据概论 | 学院 | 软件学院 软件工程专业 |
| 小组成员 | | 贺思超、陈杰、韩熔、解世超 | | |
| 课程设计题目 | | 基于往年气温数据预测未来天气气温 | | |
| 课程设计研究内容及  选题依据 | 选题依据：  天气预测一直以来就是人们的愿景以及痛点之一，有效的天气预测有利于政府对资源的统筹调度。随着社会发展，国家防灾减灾、保障社会经济发展和人民安康福祉对天气预测提出了更高的需要，发展以“数字化、精细化”为主要特征的天气预测技术是实现天气业务现代化的必然要求，大数据分析必然在其中起到关键作用。  研究内容：  基于天气预测的广泛应用和重要意义，以及其与大数据极强的关联性，本小组选择气温预测进行课程研究，旨在运用大数据分析与挖掘的理论与技术解决气温预测的相关问题。 | | | |
| 课程设计关键问题预期目标 | 由于大气运动的不确定性等因素，天气预测需要收集大量的数据用以分析处理，大数据分析与挖掘的相关理论与技术对天气预测有着重要的作用。  通过对预设时间段的天气气温分析，由最高气温、最低气温、降雨量、风力等特征值的分析进行从当前日期开始的后七天时间段的天气情况进行预测。 | | | |
| 课程设计的研究计划及预期解决方案 | 研究计划：  本实验拟采用Python语言，从www.meteomanz.com网页爬取气温数据并存储为CSV文件，使用panda读取CSV文件并进行数据预处理，并采用Sklearn框架进行机器学习处理。最后使用随机数森林模型进行建模预测。  预期解决方案：  数据来源：http://www.meteomanz.com/  存储：CSV文件  模型：随机森林模型  评估：MAE  可视化：Matplot绘制折线图 | | | |
| 技术难点及  解决方案 | 技术难点：   1. 随机森林在解决回归问题时，并没有像它在分类中表现的那么好，这是因为它并不能给出一个连续的输出。当进行回归时，随机森林不能够做出超越训练集数据范围的预测，这可能导致在某些特定噪声的数据进行建模时出现过度拟合。 2. 在现实情况中，一些数据集是不完整的或数据是文本，所以要先对数据集预处理   解决方案：   1. 增加数据集数量 2. 空数据处理根据情况用sklearn下的imputer或drop掉丢失数据的那一列。比如丢失数据情况小于80%就drop掉，反之就用imputer填充 3. 分类处理，为了把不同类型的数据转换到机器学习接受的数字类型的数据。用sklearn库下的hotlablel或者label把非数字数值分类。前者是把每一个非数值都新建一个列存放，后者是把每一个非数值标号，为了达到最高效，应该把种类数目不一样的列表分别处理，比如种类少就用hotlable，多就用label。 | | | |
| 组内分工及  开题  当前进展 | 组内分工：  解世超-数据挖掘 25%  陈杰-数据处理 25%  贺思超-数据建模 25%  韩熔-主程序构建 25%  当前进展：   1. 确立信息来源网站 2. 已选择机器学习模型 3. 已选择存储文件 4. 已确定评估以及可视化方案 | | | |