**南京中医药大学信息技术学院**

**毕业设计（论文）开题报告**

课 题 名 称：中医藏象辨证量化诊断系统的

设计与实现

学 生 姓 名： 赵壮

学 号： 084015125

班 级： 计算机科学与技术151

指 导 教 师： 杨涛

2018年 　12 月 　28　 日

**说 明**

1．根据教育部对毕业设计（论文）的评估标准，学生必须撰写《毕业设计（论文）开题报告》，由指导教师签署意见、毕业设计（论文）指导小组审查，学院教学院长批准后实施。

2．开题报告是毕业设计（论文）答辩委员会对学生答辩资格审查的依据材料之一。学生应当在毕业设计（论文）工作前期内完成，开题报告不合格者不得参加答辩。

3．毕业设计开题报告各项内容要实事求是，逐条认真填写。其中的文字表达要明确、严谨，语言通顺，外来语要同时用原文和中文表达。第一次出现缩写词，须注出全称。

4．本报告中，由学生本人撰写的对课题和研究工作的分析及描述，应不少于2000字，没有经过整理归纳，缺乏个人见解仅仅从网上下载材料拼凑而成的开题报告按不合格论。

5．开题报告原则上在第七学期18周前完成，各教研室完成毕业设计开题检查后，各设计指导小组应写一份开题情况总结报告报学院。

**南京中医药大学信息技术学院**

**毕业设计(论文)开题报告**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学生姓名 | 赵壮 | 学 号 | 084015125 | 班 级 | | 计算机151 |
| 指导教师姓名 | 杨涛 | 职 称 | 讲师 | 所在单位  或部门 | | 信息技术学院 |
| 课题来源 | 教师科研项目的全部或部分研究内容 | | | 课题类型 | | 应用基础研究 |
| 课题名称 | 中医藏象辨证量化诊断系统的设计与实现 | | | | | |
| 毕业设计的内容和意义 | **1.内容**  对深度学习、集成学习等算法进行深入研究，探索利用相关算法开展中医辨证的创新研究与应用，利用相关算法实现临床数据采集、量化诊断和可视化结果展示。具体功能如下：  a) 在藏象辨证体系指导下，利用算法实现中医智能诊断，即从临床信息到中医证型的计算；  b) 利用算法对中医医案进行分析，从准确率、召回率、F1值等比较模型优劣；多标记问题采用汉明损失、覆盖率、one-error等进行评价；  c) 设计智能辨证程序，将算法集成到程序中，实现输入“症状信息”得“辨证结果”；  d) 设计两种用户交互模式，包括单个病人智能辨证和批量数据导入后的智能辨证；数据导入至少支持1-2种格式，建议采用文本、Excel等常用格式；相关结果存储到数据库中，用于历史记录的查询；  e) 历史辨证数据的可视化结果展示，包括症状体征的分布、证型的分布等,采用图表展示。  **2.意义**  a) 本毕业设计课题尝试利用集成学习、深度学习等机器学习算法从中医临床数据中挖掘常见证型的量化诊断规律，形成“以数据为支撑”、“以算法为驱动”的辨证量化诊断体系，为大数据时代中医辨证研究提供参考；  b) 通过本毕业设计的实践，可以学习到中医藏象辨证理论，理解机器学习算法实现与优化原理，掌握Web开发技术，为进一步的科研或工作打下理论与实践基础。 | | | | | |
| 文献综述 | 本课题围绕根据中医藏象辨证理论如何利用现有的机器学习算法对中医医案、门诊病历数据进行量化诊断展开的，所以课题在理论准备阶段要通过查阅文献对以下若干问题的探讨：  a) 在中医理论的辨证论治体系中，是否存在由症状信息到证型的映射规律？  b) 现有的机器学习算法能否实现这样的映射关系？  c) 这些算法的思想、流程是什么？  d) 如何实现一个用户友好的量化诊断的系统？  本文将在查阅大量文献的基础上对以上问题进行探究与综述。  **1.从症状信息量化诊断出证型这一过程的理论可行性问题**  要解决这一问题，首先要弄清中医理论中辨证论治是一种怎样的模型。中医模型系统的基础是“天人合一”的思想，人体——这个宇宙及社会的相似物中的每一个成分都与自然、社会的相应部分相互联系着，人体内部的运转机制也同宇宙和社会运转的机制相似，并与之发生作用。因此，中医的诊断模型是将人体的内部结构和运转机制看作黑箱，对于人体内部的变化往往用对体外事物观察的结果去解释。[1]用符号化工具来描述这一模型就如下图1所示：    图1 中医对人体的识别和判断  其中W、X、Y、Z为四组内部相互之间有特定关系的直观事物，W表示对人体产生作用和影响的事物的状态；X表示正常人体内部事物的状态；Y表示病人身体内部事物的状态；Z表示病人外表事物的状态。由此看出中医保留了人类直观地认识事物的自然特征，通过一个个模型之间的转换(时间的、空间的、类比的、隐喻的、投射的、扩大或缩小的等无需习得的方式)达到对对象的了解，发展了非量化的直观推理形式。基于这样的推理中医形成了“人－症－病－证”的辨证论治体系。[2]  在这个体系里，“症”包括症状和体征，症状是主观体会到的痛苦或不适,体征是医生客观检查到的异常；“证”指证候，是疾病过程中某一阶段或某一类型的病理概括，一般由一组相对固定的、有内在联系的、能揭示疾病某一阶段或某一类型病变本质的症状和体征构成。中医临床诊断，“症”是确立“证”的主要依据，因此在中医的理论中，由症状信息到证型的映射是存在的。  **2.用机器学习的方法解决完成中医推理的技术可行性问题**  要解决这个问题，首先要解决机器学习的定义及其作用范围。机器学习是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能的人工智能分支。  机器学习算法分为四类，将实例数据划分到合适的分类中的分类算法、用于预测数值型数据的回归算法、将数据集合分成由类似的对象组成的多个类的聚类算法以及寻找描述数据统计值的密度估计算法。  显然，机器学习的分类算法能够满足本课题的任务需求，即根据症状信息将病人分到对应证型类里去。分类算法又分为传统的分类算法和深度学习算法，而传统的分类算法包括K-近邻算法、决策树算法、朴素贝叶斯、Logistic回归、支持向量机以及由弱分类器组合而成的强分类器的集成学习算法。[3]  根据李凯的研究实验中得出的结论，基于数据的集成学习算法在大多数数据集上优于基于特征集的集成学习算法且基于数据集的集成学习算法中Bagging算法和Boosting算法的表现无论是正确率还是泛化性能都优于其他算法。[4]故本课题选择Bagging算法的代表随机森林算法和Boosting算法的代表AdaBoost算法以及深度学习算法。  **3.集成学习与深度学习的算法思想和流程是怎样的**  集成学习是将多个不同的单个模型组合成一个模型，其目的是利用这些单个模型之间的差异，来改善模型的泛化性能。当学习模型具有较高的正确率且具有差异性时，则模型集成的效果明显。[5]  其中，Bagging思想是多次放回采样通过学习算法形成多个弱分类器，投票（权重相等）形成强分类器[6]，算法具体流程如下：  图2 Bagging算法流程  而Boosting算法思想是每次使用全部样本对多个弱分类器进行线性组装，各分类器根据错误率调整的权重投票强分类器。具体算法流程如下：  图3 Boosting算法流程  总结起来，Bagging和Boosting的区别如下：  a) 样本选择上：  Bagging：训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的；  Boosting：每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化，而权值是根据上一轮的分类结果进行调整。  b) 样例权重：  Bagging：使用均匀取样，每个样例的权重相等；  Boosting：根据错误率不断调整样例的权值，错误率越大则权重越大。  c) 预测函数：  Bagging：所有预测函数的权重相等；  Boosting：每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重。  d) 并行计算：  Bagging：各个预测函数可以并行生成；  Boosting：各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。[7]  深度学习的思想是基于样本数据通过一定的训练方法得到包含多个层级的深度网络结构，这个深度网络结构包含大量的单一神经元，每个神经元与大量其他神经元相连接，神经元间的连接强度在学习过程中修改并决定网络的功能。[8]  其算法的流程如下：  a) 根据学习任务即输入输出设计神经网络结构；  b) 将样本输入网络计算输出结果；  c) 根据输出与实际样本标签误差选择误差反向传播算法；  d) 调整神经元权重，训练神经网络；  e) 使用对神经网络性能进行验证，调整网络结构，重复b-e过程直到验证结果满意为止。  **4.如何实现一个用户友好的量化诊断系统**  由于本课题的涉及的大量算法开发，而Python语言对于科学计算的工具支持具有良好的生态环境，故选用现有的Python Web开发框架，其包含了多种组件，可以实现关系映射、动态内存管理、界面管理等功能。Django开发采用DRY原则，同时拥有独立的轻量级Web服务器，能快速开发Web应用。[9]  Django开发遵循MVC模式，包括模型·、视图、控制三部分：  a) 模型层是应用程序底层, 主要负责处理与数据有关的事务, 如数据存取验证等；  b) 视图层组成应用程序的业务逻辑, 负责在网页或其他类型的文档中展示数据；  c) 控制层负责从用户端收集用户的输入，可以看成提供View的反向功能。  Django的工作过程如下：  a) 浏览器向Web服务器发送HTTP请求；  b) Web服务器收到请求转发给Django应用；  c) Django请求中间件层，按照URLconf模式来匹配、映射到合适的视图函数上，视图函数利用模型和模板，按照要求生成相应的响应，中间件层把响应封装成HTTP响应，返回给Web服务器；  d) Web服务器把响应转发给客户端浏览器。[10]  具体工作过程如图4所示：    图4 Django工作过程  以上是本课题开题相关文献的综述，系统将在上述文献指导下进行开发。  参考文献  [1]王志康.回归自然的认知策略——中医诊断的模型化推理及其方法论启示[J].自然辩证法通讯,2009,31(03):26-31+25+110.  [2]江泳,陈建杉,江瑞云,郭子光.论辨证论治的完整体系:人-症-病-证[J].中医杂志,2011,52(17):1447-1450.  [3] Peter Harrington．机器学习实战[M]．北京：人民邮电出版社，2013：3-9．  [4]李凯,崔丽娟.集成学习算法的差异性及性能比较[J].计算机工程,2008(06):35-37.  [5]付忠良,赵向辉.分类器动态组合及基于分类器组合的集成学习算法[J].四川大学学报(工程科学版),2011,43(02):58-65.  [6]徐继伟,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(06):1082-1092.  [7]李晓波.集成分类对比:Bagging NB & Boosting NB[J].微电子学与计算机,2010,27(08):136-139.  [8]尹宝才,王文通,王立春.深度学习研究综述[J].北京工业大学学报,2015,41(01):48-59.  [9]朱贇.Python语言的Web开发应用[J].电脑知识与技术,2017,13(32):95-96.  [10]龚新定,余艳梅,吴小强,何小海.基于Django的实验室信息管理系统设计[J].微型机与应用,2016,35(22):108-111. | | | | | |
| 设计方案和技术路线 | **1.设计方案**  本系统采用B/S架构，使用Python语言进行后端开发，Html、Javascript、CSS进行前端开发。系统分为五个功能模块，即数据采集模块、Web交互模块、数据处理模块、数据存储模块以及数据展示模块：  a) 数据采集模块，顾名思义，用于收集用户门诊信息，支持单项表单录入或语音输入，以及批量文件导入。表单录入用Javascript进行预处理之后发送给服务器进行进一步处理；语音输入利用云服务提供的语音识别接口将音频文件转化为文字信息，再由用户编辑确认发送给服务器处理；文件批量导入功能利用Python强大的工具库对用户本地进行读写操作，提取多条患者信息；  b) Web交互模块，即负责与用户进行交互，同时协调后端组件进行工作，系统基于Python强大的Django框架进行开发，采用MVC的软件设计模式，控制器负责接受处理用户输入请求，指示响应操作，模型负责与数据库以及数据处理模块的交互，对数据进行读写并发送给视图，视图负责处理要返回给用户的界面；  c) 数据处理模块，即负责算法模型对数据的处理，利用Sklearn包的AdaBoost、随机森林等算法进行集成学习，利用Tensorflow框架进行深度学习，对输入的症状、体征信息进行量化诊断并返回准确率、召回率等指标，以便进行优化调整；  d) 数据存储模块，本系统采用MySQL数据库，仅需要MySQLdb库即可用Python与之交互读写；  e) 数据展示模块，即负责向用户直观地展示历史诊断结果，系统采用Echarts.js和D3.js脚本进行作图。  具体设计方案如下图5所示：  图5 系统的设计方案  **2.技术路线**    图6 系统的技术路线 | | | | | |
| 设计进度 | 2月20日——2月29日：前后端设计与开发  3月01日——3月15日：机器学习算法设计与编码实现  3月15日——3月31日：模型训练与优化  4月01日——4月15日：项目完善与改进  4月15日——4月30日：论文撰写  5月01日——5月15日：准备答辩材料  5月15日——6月20日：课题扩展开发 | | | | | |
| 特色与创新 | **1.特色**  a) 项目“中医藏象辨证体系”为理论基础，将各种中医辨证方法科学地融为一体，利用病证结合的思维和方法，“以象测藏，从症辨证”，设计了量化诊断模型；  b) 利用机器学习中效果出众的集成学习与深度学习算法对中医医案进行辨证规律的探索，从准确率、召回率、F1值等参数为评价模型性能提供参考，满足科研需求；  c) 利用数据可视化技术，将症状体征的分布、证型的分布等历史辨证数据直观展现，更有利于量化诊断中潜在的辨证规律的探索与发现。  **2.创新**  a) 在方法上突破创新，将用于图像分类的卷积神经网络模型对医案这种文本数据进行分类，考虑症状信息在中医理论中的具体意义与关系，对特征分布进行设计与尝试；  b) 将算法与临床实践经验结合，为算法适配更具有实际参考意义的参数，提高算法性能与效果。 | | | | | |
| 指导教师  意 见 | 指导教师签名：  年 月 日 | | 毕业设计指导小组意见 | | 组长(签名)：  年 月 日 | |
| 学院意见 | 教学院长(签名)：  年 月 日 | | | | | |