**赛尔网络下一代互联网技术创新项目**

**结题技术报告**

|  |  |
| --- | --- |
| 项目编号： | NGII20170703 |
| 项目名称： | IPv6环境下支持隐私保护的医疗大数据挖掘方法研究 |
| 技术领域： | IPv6智能医疗 |
| 承担单位： | 陕西师范大学 （盖章） |
| 项目负责人： | 王寒 |
| 指导教师： | 李蜀瑜 |
| 项目期限： | 一年 |
| 填报日期： | 2019年 05 月 12 日 |
| 联系人： | 陈竑毓 |
| 联系电话： | 18072210523 |

**中国教育和科研计算机网CERNET网络中心制**

**二〇一七年五月填 写 说 明**

一、请严格按照表中要求填写各项，用A4纸打印（单双均可），按照要求逐一签字盖章。

二、本报告的“项目名称”和“项目编号”等项目基本信息应与《任务合同书》一致；已做调整的按照调整批准后的信息填写。

三、本报告第一次出现外文名词时，要写清全称和缩写，再出现同一词时可以使用缩写。

四、编写人员应客观、真实地填报，尊重他人知识产权，遵守国家有关知识产权法规。

**项目基本信息**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | | | IPv6环境下支持隐私保护的医疗大数据挖掘方法研究 | | | |
| 技术领域 | | | IPv6智能医疗 | | | |
| 项目期限 | | | 一年 □两年 | | | |
| 成果类型 | | | □硬件 □软件平台 软件应用系统 □APP □网站 专利 □技术标准 开发源码 软件著作权 论文论著 □其他 | | | |
| 申请  单位  信息 | | 单位名称 | | 陕西师范大学 | | |
| 通讯地址 | | 西安市长安区西长安街620号 | 邮政编码 | 710119 |
| 所在地区 | | 陕西省西安市 | 组织机构代码或统一社会信用代码 | 43523253-x |
| 校科技主管部门 | | 陕西师范大学科学技术处 | 联系电话 | 029-85310340 |
| 电子信箱 | | kjc@snnu.edu.cn | 传真号码 | 029-85310230 |
| 项目  申请人  信息 | 学生 | 姓名 | | 王寒 | 性别 | 男 |
| 出生日期 | | 1994-01-01 | 年级 | 2016级硕士研究生 |
| 所学专业 | | 软件工程 | 所在院系 | 计算机科学学院 |
| 固定电话 | | 029-85310552 | 移动电话 | 13892900811 |
| 传真号码 | | 029-85310554 | 电子信箱 | wanghan061@163.com |
| 证件类型 | | 身份证 | 证件号码 | 610481199401010037 |
| 指导教师 | 姓名 | | 李蜀瑜 | 性别 | 男 |
| 年龄 | | 41 | 职称/学位 | 副教授**/**博士 |
| 专业领域 | | 计算机软件 | 所在院系/部门 | 计算机科学学院**/**网络信息中心 |
| 移动电话 | | 13572128758 | 固定电话 | 029-85310552 |
| 传真号码 | | 029-85310554 | 电子邮箱 | lishuyu@snnu.edu.cn |
| 证件类型 | | 身份证 | 证件号码 | 610302197801202011 |
| 资助经费（万元） | | | | | 10.00 | |

# 项目研究内容（请按照实际研究工作填写）

1. 面向医疗大数据挖掘的隐私保护方法研究

针对医疗大数据挖掘过程中可能存在的隐私泄露风险，分析典型的隐私保护方法的优缺点，从理论上研究差分隐私模型与数据挖掘方法的结合，并以癫痫疾病发病预测和心脏病、帕金森的患病预测作为应用场景研究支持隐私保护的医疗大数据分类和聚类方法。

1. 基于差分隐私和深度学习的医疗大数据分类算法研究

针对癫痫疾病的发病预测需求，提出一种发病状态分类算法，采用基于深度学习长短时记忆模型的分类方法进行发病状态分类，在此之上，通过改进模型的损失函数，向梯度值中加入正则项以提高模型泛化能力，并向模型更新过程中的计算的平均梯度值添加噪声实现对训练数据集的差分隐私保护。

1. 基于差分隐私和小波聚类的医疗大数据聚类算法研究

针对心脏病和帕金森病的患病预测需求，提出一种患病状态聚类算法，结合差分隐私模型采用小波聚类算法进行患病状态聚类，在此之上，通过在量化步骤和网格识别步骤中添加噪声，对加噪后的数据进行小波变换并利用广度优先搜索实现支持隐私保护的聚类算法。

1. IPv6环境下支持隐私保护的医疗大数据挖掘系统实现

采用TensorFlow深度学习框架实现基于差分隐私和深度学习的医疗大数据分类算法，采用网格映射技术实现基于差分隐私和小波聚类的医疗大数据聚类算法。基于IPv6网络环境，采用Java Web技术搭建疾病预测系统，利用上述算法设计实现疾病预测模块并应用于系统中，通过对该系统的部署和实际运行来验证研究成果的有效性。

# 项目研究的技术路线与方法（请按照实际研究工作填写）

**研究方法**

采用理论与应用相结合的方式，在模型构建与算法设计的基础上，实现一个支持隐私保护的医疗大数据挖掘系统，通过对真实医疗数据的挖掘，从应用角度验证所设计算法的实用性。

**技术路线**

（1）面向医疗大数据挖掘的隐私保护方法研究

针对医疗大数据的隐私保护问题，近年来出现了不少基于匿名化和划分等技术的传统隐私保护方法。但这些方法，总是因新型攻击的出现而需要不断完善，此外在面对海量数据时，往往无法达到预期的隐私保护效果。差分隐私模型作为一种严格的和可证明的隐私保护方法，它不关心攻击者的背景知识，并且提供了隐私保护的量化评估能力。基于上述事实，本项目采用差分隐私模型实现支持隐私保护的医疗大数据挖掘。

差分隐私的形式化定义如下：

 公式（1）

其中参数称为隐私保护预算，与为一对邻近数据集。

目前，在基于差分隐私的数据发布和数据挖掘方向，多数研究工作偏重于发布算法和挖掘算法本身，对与医疗大数据领域的融合研究较少。此外，将相关算法的理论研究与具体实现相结合的系统也相对较少。本项目针对医疗大数据，在理论分析和算法设计的基础上，通过将算法应用于实际系统的方式实现支持隐私保护的医疗大数据挖掘方法的研究工作。

（2）基于差分隐私和深度学习的医疗大数据分类算法研究

基于神经网络的机器学习技术广泛应用在多个领域，其自适应信息的处理能力在分类方面表现尤为突出。例如，在医疗领域，一方面，对糖尿病患者的医疗信息按照年龄段进行分类，可得到各个年龄段的患者数量，以此来鉴别糖尿病的高发群体及其特征，从而有效地辅助医生的治疗。但另一方面，一旦患者的隐私数据泄露，将会造成非常严重的后果。因此，本项目针对癫痫疾病的发病预测需求，提出一种基于差分隐私和深度学习的分类算法，期望在实现隐私保护的同时，保证发病状态分类的准确性。

算法采用深度学习中的长短时记忆模型（Long-Short Term Memory, LSTM），该模型是一种有效的链式循环神经网络。通过改进LSTM模型的损失函数，向梯度值中加入正则项来提高模型泛化能力并求其平均值，并通过对模型中的平均梯度添加噪声来实现支持隐私保护的分类算法。

分类的最终结果通过损失函数来进行衡量，损失函数的定义如下：

 （公式2）

其中，*m*是样本数量，**是模型参数（权值和偏置），*x*是每个样本i的特征，*y*是期望值，为网络输出的预测值，为范数，为正则项系数。

采用随机梯度下降（stochastic gradient descent, SGD）算法来完成损失函数最小化，按参数的更新梯度负方向，来更新每个，其中为学习率：

 （公式3）

针对分类过程中可能存在的隐私泄露风险，提出一种基于差分隐私和深度学习的分类算法。针对LSTM模型的损失函数进行改进并加入噪声。损失函数公式更新为：

 （公式4）

其中，是函数的敏感度，是标准差，是一个均值为0，标准差为的（高斯）正态分布。

医疗数据按批次进入网络后，在使用SGD算法进行梯度更新时，根据损失函数计算出梯度，并在其中加入正则项来提高模型泛化能力，计算其平均值并加入噪声形成新的损失函数，通过参数更新最终得到稳定的网络模型，在实现医疗数据分类的同时满足隐私保护。

支持隐私保护的分类算法流程如图1所示。

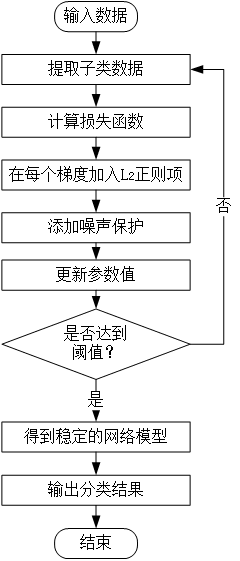


图1 基于差分隐私和深度学习的分类算法流程

（3）基于差分隐私和小波聚类的医疗大数据聚类算法研究

针对医疗大数据的聚类分析，本文拟采用基于网格的小波聚类算法。小波聚类算法能有效发现任意形状的簇，且可广泛地应用于大数据集的处理。例如，对癫痫脑电图数据进行小波聚类得到分组结果，通过对组内的共同特征与患者的关联性分析，便于医生掌握当前患者状态从而进行诊断。

针对聚类过程中可能存在隐私泄露的风险，本项目拟提出一种基于差分隐私的小波聚类算法。由于高维空间数据稀疏，量化形成的矩阵存在大量的空单元，考虑采用元胞数组存储方式，保留组内非空元素并加入拉普拉斯噪声进行保护，既减少噪音的影响又节约空间开销。对加噪后的数据进行小波变换，得到特征数据形成的序列组。提取特征序列组，采用广度优先搜索寻找连通单元并形成簇。针对网格映射所形成的簇序号，依据查找表即可判断出数据对象所属的类，最终完成聚类。

查找表的映射关系如图2所示。

图2 查找表映射关系图

这里的小波变换拟采用离散型。小波变换公式如下：

 （公式5）

为离散小波函数，定义如下：

 （公式6）

其中a和b分别为尺度因子和时移因子，。

设查询函数，量化后的数据集，查询结果为，为敏感度，通过在上加入拉普拉斯噪声来保护隐私，函数响应值为：

 （公式7）

针对小波变换形成的序列，如果直接舍弃序列中的非正值，会造成聚类结果严重失真，但如果直接发布非正值序列则可能会造成隐私泄露。本项目拟采用对非正值计数并加噪的方式，保证算法满足差分隐私定义。通过加入随机噪声，序列中部分非正值变成正值，因此对于原有正值序列，舍弃掉部分正值，以提高算法实用性。

支持隐私保护的聚类算法流程如图3所示。

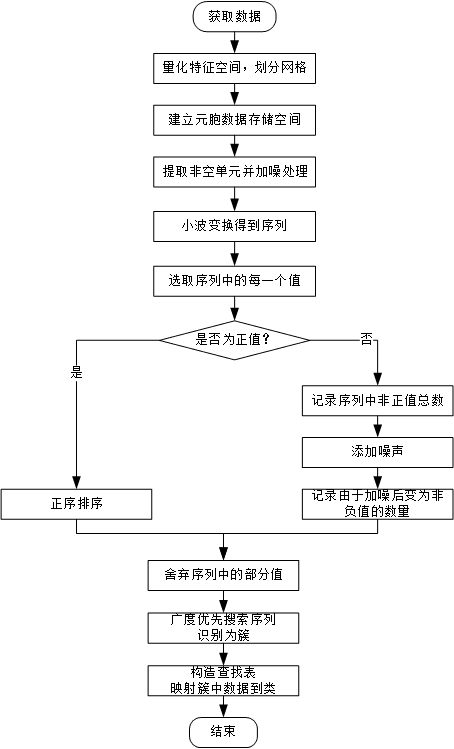


图3 基于差分隐私的小波聚类算法流程

（4）IPv6环境下支持隐私保护的医疗大数据挖掘系统实现

在算法具体实现上，医疗大数据分类算法选用UCI的Epileptic Seizure Recognition（癫痫发作识别）数据集用于训练LSTM模型进行癫痫发作状态的识别分类，其中Epileptic Seizure Recognition数据集是常用的癫痫发作检测的时间序列数据集，包含11500个实例，179个属性，详细记录了500个人23.6秒内的脑电图数据，相应的时间序列采样到4097个数据点，每个数据点都是脑电数据记录的值，并具有五个发作状态标签，实际应用中将其处理成二分类问题。医疗大数据聚类算法选用UCI的心脏病（Heart Disease Dataset）和帕金森病（Parkinson's Disease）两种疾病数据集为测试数据集，其中心脏病数据集实例数303，包含静息血压、血清胆固醇、空腹血糖、吸烟年数、静息心电图、运动心电图等属性，根据这些属性对患者不同年龄、不同性别的患者数据进行聚类分析，将患者归属于非心痛类型、典型心绞痛、非典型心绞痛、非心绞痛、无症状等五种类型。帕金森病数据集包含756个实例，通过此数据聚类分析，预测患者是否患有帕金森病。

在应用系统实现上，基于IPv6网络环境，采用Java Web技术设计实现一个支持隐私保护的医疗大数据挖掘系统，将上述提出的分类算法和聚类算法应用在挖掘模块中，通过对实际医疗大数据的挖掘来验证所提算法的可行性与实用性。

# 项目取得的研究成果（请对项目取得的主要研究成果进行详细说明）

**软件成果**

一个基于B/S架构的IPv6环境下支持隐私保护的疾病预测系统，该系统由账户管理、留言管理、管理平台、数据平台、预警平台以及医患社区等六个功能模块组成。

其中账户管理模块的主要功能包括：用户注册、用户删除、用户登录、用户信息添加、用户信息修改、角色与权限管理等。通过此功能模块可以实现患者、医生、管理员等三种角色权限管理以及信息编辑等。

留言管理模块的主要功能包括：留言添加、留言处理状态修改以及留言删除等。通过此功能模块不同角色可以对留言信息编辑等。

管理平台模块的主要功能包括：患者挂号、修改挂号信息、取消挂号、开具诊断记录、查看既往病史、查看就诊记录等。通过此功能模块可以实现在线医疗办公，提高医疗看病效率。

数据平台模块的主要功能包括：共享文件管理、个人文件管理、模型训练、疾病预测可视化等。通过此功能模型用户可以便捷的将个人文件存储在云端，实现医疗数据的管理。医生还可以通过模型训练病人数据，实现病人疾病预测可视化，辅助医疗，提高诊断准确率。

预警平台模块的主要功能包括：发病总览、实时数据可视化、实时发病记录诊断、异常记录信息补充、异常记录删除、统计报告可视化等。通过此功能模块医生可以更加详细了解患者信息，提前预防疾病发生，有效提高医疗诊断水平。

医患社区模块的主要功能包括：文章列表、模糊筛选、评论文章、新增文章、修改文章、删除文章等。通过此功能模型，实现了患者与医生之间的互动交流，医疗知识的共享。

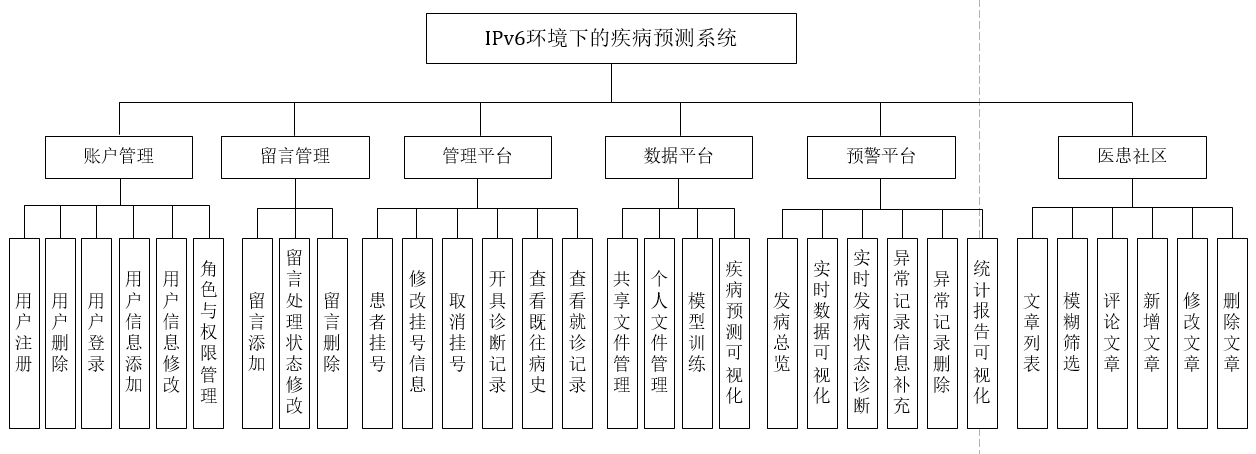


图4 疾病预测系统功能模块图

**在国内学术期刊和国际会议上发表论文**

1. Li Y, Li S. A Real-Time Location Privacy Protection Method Based on Space Transformation[C]//2018 14th International Conference on Computational Intelligence and Security (CIS). IEEE, 2018: 291-295. （EI检索）
2. Li Y, Li S. Research on Differential Private Streaming Histogram Publication Algorithm[C]//2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2019: 598-603.
3. Wang H, Li S. Differential Private Multiple Classification Algorithm for SVM[C]//2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2019: 604-609.
4. Wang H, Li S. Research on Social Networks Publishing Method under Differential Privacy[C]//2019 2nd EAI International Conference on Security and Privacy in New Computing Environments(SPNCE).Springer
5. Zhang G, Li S. Research on Differentially Private Bayesian Classification Algorithm for Data Streams[C]// 2019 the 4th IEEE International Conference on Big Data Analytics (ICBDA 2019). IEEE
6. Bian J, Li S. Research on a Privacy Preserving Clustering Method for Social Network[C]// 2019 the 4th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA 2019). IEEE

**已申请专利**

1. 2018年11月 发明创造名称：一种基于哈希前缀树的数据完整性验证方法 申请号：201811351086.7 申请人：陕西师范大学
2. 2019年02月 发明创造名称：基于贝叶斯网络的民航突发事件因果关系分析方法 申请号：201910132202.4 申请人：陕西师范大学、赛尔网络有限公司

**已申请软件著作权**

1. 2018年03月 软件名称：支持PDP协议的健康监护系统V1.0 登记号：2018SR870116 著作权人：陕西师范大学、赛尔网络有限公司
2. 2018年10月 软件名称：基于IPv6和区块链的农产品溯源系统[简称：农产品溯源系统]V1.0 登记号：2018SR870119 著作权人：陕西师范大学、赛尔网络有限公司
3. 2018年03月 软件名称：关于构件的应用规划与部署工具软件V1.0 登记号：2018SR193135 著作权人：陕西师范大学
4. 2018年08月 软件名称：齿轮箱实时诊断系统V1.0 登记号：2018SR680518著作权人：陕西师范大学
5. 2018年05月 软件名称：民用机载软件测试平台V1.0 登记号：2018SR680559 著作权人：陕西师范大学

# 项目研究成果的市场推广应用前景分析

我国庞大的人口导致医疗需求巨大，同时随着老龄化现象加剧，不健康的生活方式导致慢性疾病多发于人群，加剧了医疗资源的短缺。而物联网、通信技术等的快速发展，推进了医疗行业的信息化、数据化，同时也催生了医疗大数据的发展。医疗大数据作为国家重要的发展战略，受到市场认可、资本扶持，其通过建立以患者为中心的医疗体系，对解决医疗资源短缺、不均衡等问题具有重大作用。

利用医疗大数据解决问题的同时，其带来的隐私问题也不容小觑，能否在提供医疗便利的同时保护用户隐私面临着极大的挑战。差分隐私作为一种较强的隐私保护方法，其对攻击模式是不敏感的，能够很好地应用于数据分析过程。同时，我国正积极推进IPv6网络的建设，构建下一代互联网产业生态体系。IPv6相较IPv4，具有更大的地址空间，更快的传输效率，以及更高的安全性。

本项目基于IPv6网络环境，设计实现支持隐私保护的疾病预测系统，适应市场需求，响应国家发展战略，具有良好的市场前景。

# 项目存在问题、建议及下一步设想

1. 算法在提供隐私保护的同时，会降低其效用，在下一步的工作中，计划改进差分隐私结合方式，实现在同等隐私保护程度下，提升数据效用。
2. 系统当前支持的用于癫痫发病预测的深度学习算法的模型结构及优化方法等已经固定，在下一步的工作中，拟提供多种学习模型、优化方法、激活函数等供用户自主选择，通过构建不同模型结构能实现对更多疾病的支持，以及提高预测准确度。

# 相关人员与部门意见

项目负责人（签字）：

指导教师（签字）：

主管部门（负责人签字或部门签章）：

年 月 日

# 承担单位意见

单位盖章：

年 月 日