心血管疾病预测中的多标签分类算法研究与并行化实现

基于多标签分类的心血管疾病预测模型研究与应用

心血管疾病预测多标签分类模型研究与应用

论文进行的主要工作

（1）

（2）预测目标分析

（3）模型设计

（4）数据预处理、的设计实现

1）相关研究（国内外）

2）心血管疾病预测目标

数据--------分类结果（预处理-特征 ）

3）数据

本文以心血管疾病预测模型建立为主，在医学基础上利用多标签技术准确预测出患者早期可能存在的多种疾病，以达到早治疗早康复的目的。建模过程中解决了心血管病影响因子选择、基于多标签的数据不平衡问题、多标签应用建模等多种问题，主要工作如下：

  1.根据医学规则，利用spark平台和ETL技术，从千万条医疗数据记录中抽取心血管疾病人群作为实验组，相应的健康人群作为对照组，由于医疗记录存在数据缺失量大、数据异常等问题，需要对数据进行异常值检测、缺失值处理、属性转换等操作。

  2.为了预测结果的准确性、可靠性，对数据进行降维处理，通过统计检验和分类算法，提取影响心血管疾病发展的有效因子，具体分两步：

 一是 对单个心血管病进行特征相关性分析，通过spearmanr检验，去除共线性强的特征，影响目标疾病的有效因子提取出来。

  二是 应用逻辑斯蒂回归分类算法分别评估单个心血管病的分类效果，根据特征权重排序，选择权重较高的特征作为该病的度量指标，最终合并所有心血管病的特征，保留效果最好的特征集合。

3.由于医疗数据分布的稀疏性、不平衡性，使用多标签统计评价数据集的方法对目标标签集进行统计分析，评估该标签集是否有多标签分类的意义，根据结果，我采用了过采样、标签删除合并策略，将相关性强的标签且量少的标签合并，同时尽可能的保留更多的标签间的信息，多次迭代最终多标签数据统计信息中Cardinality为2.24，平衡了标签集分布。

  4. 应用多标签分类算法评价在当前数据集上的表现。约10000多个样本,按70%划分训练集测试集，利用BR-logisticRegression、AdaBoostMH-logisticRegression、MLKNN等多种多标签算法进行分类，模型结果具有可行性。

根据题目“心血管疾病预测中的多标签分类算法研究与并行化实现”论文的工作主线希望为：

1）采用的心血管疾病预测模型是什么？

2）预测模型采用的数据集特性？

3）多标签统计评价数据集方法研究（已有方法）

4）针对项目特定数据集的多标签统计评价数据集方法研究

5）方法的实现

6）方法的测试与评价。