

硕士学位论文

**基于多标签分类的心血管疾病预测模型研究与应用**

学位申请人：程敬

指导教师：魏恒义高工

类别（领域）：计算机技术

2018年6月

**Research and Application of Prediction Model of Cardiovascular disease based on Multi-label Classification**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Jing Cheng

Supervisor: S.E.Hengyi Wei

(Computer Technology)

June 2018

**论文题目：基于多标签分类的心血管疾病预测模型研究与应用**

**类别（领域）：计算机技术**

**学位申请人：程敬**

**指导教师：魏恒义 高工**

摘 要

心血管疾病是一种严重威胁人类的常见病，具有高患病率、高致残率和高死亡率的特点，所以做到早预测，早治疗，提高心血管病患者的生存率显得极其重要。机器学习领域的多标签分类具有单样本多语义识别功能，可有效预测一个病人可能存在的多种并发症，一种基因多种功能等问题，为解决复杂的医学问题提供了有力手段。

本文以实际临床数据为研究的数据来源，对心血管疾病问题、多标签学习技术进行广泛深入调研的基础上，第一步认识了解数据例如数据属性代表的意义、一些常见的心血管病等，分析统计多种心血管疾病数据分布，确定研究对象，第二步对研究对象进行数据提取与预处理，由于数据量达到千万级，利用spark大数据平台完成数据加载预处理，获得多标签学习对应的数据集，第三步特征选择，采用统计学方法和分类算法进行二次特征选择，选择表现效果最好的特征集。第四步为解决心血管疾病数据集分布的稀疏性、不均衡性，针对心血管疾病数据集中大样本过度冗余、小样本缺乏数据表示等不均衡问题，提出了多标签双重自适应随机采样算法。第五步根据心血管数据集的特性，提出了一种基于混合策略的多标签分类框架来构建心血管疾病预测模型。

本文解决海量医疗数据加载与预处理、心血管疾病影响因子选择、多标签数据集不均衡性、多标签分类算法应用于心血管疾病预测与评估等问题，重点在于提出的多标签双重自适应随机采样算法，均衡了多标签数据集的分布，提高模型预测的可靠性、准确性，进而在此基础上结合数据局部性特点和全局标签间相关性特点的优势，提出了基于混合策略的多标签分类框架来构建心血管疾病预测模型，获得了好的预测性能，展示了多标签分类算法在心血管疾病预测领域的优势。

**关 键 词**：心血管疾病；多标签；多标签不均衡性；重采样；混合策略

**论文类型**：应用研究

**Title:** **Research and Application of Prediction Model of Cardiovascular disease based on Multi-label Classification**

**Professional Fields: the degree of Master of Engineering**

**Applicant: Jing Cheng**

**Supervisor: S.E.Hengyi Wei**

ABSTRACT

Cardiovascular disease is a common disease which seriously threatens human beings. It has the characteristics of high morbidity, high disability and high mortality. Therefore, early prediction, early treatment and improving the survival rate of patients with cardiovascular disease are extremely important. Multi-label classification in machine learning field has the function of recognizing multi-semantic for single sample. It can effectively predict a variety of complications of a patient and multiple functions of a gene, which provides powerful means for solving complex medical problems.

Based on the actual clinical data for the study of the data source, the paper made an extensive and profound research on the issue of cardiovascular disease and multi-label classification algorithm. The first step understood the meaning of data such as data attribute and some of the most common cardiovascular disease. Analyzing the data distribution. determining the object of study. The second steps extracted and pretreat data for the object of study. Owing to the data reached ten million, spark which is a big data platform complete loading and preprocessing for data to obtain the corresponding multi-label learning data set. Third step is feature selection, combining statistical methods with classification algorithms to select the best feature set. The fourth step is to solve the problem of sparsity and imbalance for multi-label data sets. Aiming at the imbalance problem of multi label data set, such as large sample redundancy, small sample lacking of data representation, a multi-label double adaptive random sampling algorithm is proposed. According to the characteristics of the cardiovascular data set, The fifth step is to build a multi label classification model based on hybrid strategy and apply it to predict cardiovascular disease.

In this paper, we solve these problems, such as the massive medical data loading, cardiovascular disease pretreatment, factor selection, highly unbalanced for multi-label data set, and multi-label classification algorithm applied to the prediction and evaluation of cardiovascular disease. The paper emphasizes the multi-label dual adaptive random sampling method, because the method balanced the distribution of multi-label data sets and improved reliability and accuracy model. Based on that, the advantages of data locality and global correlation between tags are also discussed. Then a multi label classification framework based on hybrid strategy is proposed to build cardiovascular disease prediction model and achieve good prediction performance.

**KEY WORDS**:Cardiovascular disease;Multi-label classification;Multi-label class imbalance; Resampling; Hybrid scheme

**TYPE OF THESIS**: Application Research

# 

目 录

[1 绪论 1](#_Toc511756893)

[1.1 课题研究背景和意义 1](#_Toc511756894)

[1.2 国内外研究现状 2](#_Toc511756895)

[1.2.1 机器学习在生物医学领域的发展现状 2](#_Toc511756896)

[1.2.2 多标签学习的研究现状 3](#_Toc511756897)

[1.3 论文主要研究内容 3](#_Toc511756898)

[1.4 论文的组织结构 4](#_Toc511756899)

[2 多标签分类相关研究 6](#_Toc511756900)

[2.1 单标签分类的局限性 6](#_Toc511756901)

[2.2 多标签学习方法 6](#_Toc511756902)

[2.2.1 多标签学习的形式化描述 6](#_Toc511756903)

[2.2.2 多标签学习的统计评价方法 7](#_Toc511756904)

[2.2.3 多标签学习方法 9](#_Toc511756905)

[2.3 标签不均衡性处理 12](#_Toc511756906)

[2.3.1 单标签分类中类别不均衡性问题 12](#_Toc511756907)

[2.3.2 多标签类别不均衡性探讨 12](#_Toc511756908)

[2.3.3 多标签类别不均衡性的问题处理方法 13](#_Toc511756909)

[2.4 本章小结 14](#_Toc511756910)

[3 心血管病预测模型研究 15](#_Toc511756911)

[3.1 模型建立的目标和步骤 15](#_Toc511756912)

[3.2 心血管疾病数据采集 16](#_Toc511756913)

[3.2.1 数据集分析 16](#_Toc511756914)

[3.2.2 研究对象及预测目标的选取 17](#_Toc511756915)

[3.2.3 数据采集实现 18](#_Toc511756916)

[3.3 患病记录数据预处理 20](#_Toc511756917)

[3.3.1 数据清洗与有效数据提取 20](#_Toc511756918)

[3.3.2 异常值处理 21](#_Toc511756919)

[3.3.3 缺失值处理 21](#_Toc511756920)

[3.3.4 数据预处理实现 22](#_Toc511756921)

[3.4 心血管疾病数据特征选择策略 24](#_Toc511756922)

[3.4.1 特征数据提取 24](#_Toc511756923)

[3.4.2 特征数据的二次选择 24](#_Toc511756924)

[3.4.3 特征选择实现 25](#_Toc511756925)

[3.5 本章小结 26](#_Toc511756926)

[4 多标签双重自适应随机采样算法 27](#_Toc511756927)

[4.1 引言 27](#_Toc511756928)

[4.2 多标签重采样算法 28](#_Toc511756929)

[4.3 多标签双重自适应采样算法 30](#_Toc511756930)

[4.3.1 算法概述 30](#_Toc511756931)

[4.3.2 采样算法选取 31](#_Toc511756932)

[4.3.3 标签集合的划分 31](#_Toc511756933)

[4.3.4 数据集均衡性标准 32](#_Toc511756934)

[4.3.5 ML-DARS算法描述 32](#_Toc511756935)

[4.4 ML-DARS算法实验 34](#_Toc511756936)

[4.4.1 实验数据集 34](#_Toc511756937)

[4.4.2 实验设置 34](#_Toc511756938)

[4.4.3 实验结果分析 35](#_Toc511756939)

[4.5 本章小结 37](#_Toc511756940)

[5 基于混合策略的心血管疾病预测模型 39](#_Toc511756941)

[5.1 引言 39](#_Toc511756942)

[5.2 基于混合策略的心血管疾病预测模型 39](#_Toc511756943)

[5.2.1 大规模数据分批训练策略 39](#_Toc511756944)

[5.2.2 模型构建思想 40](#_Toc511756945)

[5.3 实验 42](#_Toc511756946)

[5.3.1 数据集分析与处理 42](#_Toc511756947)

[5.3.2 RAKEL算法分批训练策略 44](#_Toc511756948)

[5.3.3 分类结果 44](#_Toc511756949)

[5.4 本章小结 48](#_Toc511756950)

[6 结论与展望 49](#_Toc511756951)

[6.1 总结 49](#_Toc511756952)

[6.2 展望 49](#_Toc511756953)

[致 谢 50](#_Toc511756954)

[参考文献 51](#_Toc511756955)

[攻读学位期间取得的研究成果 53](#_Toc511756956)

声明

CONTENTS

[1 Preface 1](#_Toc511749984)

[1.1 Background and Significance 1](#_Toc511749985)

[1.2 Glance of Current Research Status 2](#_Toc511749986)

[1.2.1 The development of machine learning in the biomedical field 2](#_Toc511749987)

[1.2.2 The development of multi-label learning 3](#_Toc511749988)

[1.3 Main research contents 3](#_Toc511749989)

[1.4 The structure of the paper 4](#_Toc511749990)

[2 Multilabel class imbalance and classification methods 6](#_Toc511749991)

[2.1 The limitations of single label classification 6](#_Toc511749992)

[2.2 Multi-label learning 6](#_Toc511749993)

[2.2.1 Formal description of multi label learning 6](#_Toc511749994)

[2.2.2 The statistical evaluation method of multi label learning 7](#_Toc511749995)

[2.2.3 Typical processing methods for multi label learning 9](#_Toc511749996)

[2.3 Class imbalance 12](#_Toc511749997)

[2.3.1 Class imbalance in traditional single label 12](#_Toc511749998)

[2.3.2 Class imbalance in multi label 12](#_Toc511749999)

[2.3.3 The solution of multi-label class imbalance 13](#_Toc511750000)

[2.4 Summary 14](#_Toc511750001)

[3 Prediction model of cardiovascular disease 15](#_Toc511750002)

[3.1 The goals and steps of the Model 15](#_Toc511750003)

[3.2 Data collection of cardiovascular disease 16](#_Toc511750004)

[3.2.1 Data Analysis 16](#_Toc511750005)

[3.2.2 The selection of the research object and the forecast target 17](#_Toc511750006)

[3.2.3 Implementation of Data acquisition 18](#_Toc511750007)

[3.3 Data preprocessing of cardiovascular disease 20](#_Toc511750008)

[3.3.1 Data cleaning and effective data extraction 20](#_Toc511750009)

[3.3.2 Outlier processing 21](#_Toc511750010)

[3.3.3 Missing value processing 21](#_Toc511750011)

[3.3.4 Implementation of Data preprocessing 22](#_Toc511750012)

[3.4 Feature selection strategy 24](#_Toc511750013)

[3.4.1 Feature data extraction 24](#_Toc511750014)

[3.4.2 Two selection of feature data 24](#_Toc511750015)

[3.4.3 Implementation of Feature selection 25](#_Toc511750016)

[3.5 The chapter of summary 26](#_Toc511750017)

[4 Double and adaptive random sampling algorithm in multi label 27](#_Toc511750018)

[4.1 Introduction 27](#_Toc511750019)

[4.2 Multi label resampling algorithm 28](#_Toc511750020)

[4.3 Double and adaptive random sampling algorithm in multi label 30](#_Toc511750021)

[4.3.1 An overview of the algorithm 30](#_Toc511750022)

[4.3.2 Sampling algorithm selection 31](#_Toc511750023)

[4.3.3 Partition of label set 31](#_Toc511750024)

[4.3.4 Standard of data set equilibrium 32](#_Toc511750025)

[4.3.5 Description of ML-DARS algorithm 32](#_Toc511750026)

[4.4 Experiment 34](#_Toc511750027)

[4.4.1 Dataset 34](#_Toc511750028)

[4.4.2 Setting 34](#_Toc511750029)

[4.4.3 Analysis of the results 35](#_Toc511750030)

[4.5 The chapter of Summary 37](#_Toc511750031)

[5 Prediction model of cardiovascular disease based on mixed strategy 39](#_Toc511750032)

[5.1 Introduction 39](#_Toc511750033)

[5.2 Prediction model of cardiovascular disease based on mixed strategy 39](#_Toc511750034)

[5.2.1 Large scale data batch training strategy 39](#_Toc511750035)

[5.2.2 Model building thought 40](#_Toc511750036)

[5.3 Experiment 42](#_Toc511750037)

[5.3.1 Data set analysis and processing 42](#_Toc511750038)

[5.3.2 Batch training strategy of RAKEL algorithm 44](#_Toc511750039)

[5.3.3 Classification results 44](#_Toc511750040)

[5.4 Summary 48](#_Toc511750041)

[6 Conclusion and prospect 49](#_Toc511750042)

[6.1 Conclusion 49](#_Toc511750043)

[6.2 Prospect 49](#_Toc511750044)

Acknowledgements [50](#_Toc511750045)

[References 51](#_Toc511750046)

Achievements [53](#_Toc511750047)

Declarations

# 绪论

## 课题研究背景和意义

近年来，我国医疗体制改革不断深化，医疗领域的信息化不断完善，每天产生出大量的医疗数据，这些数据中不仅包括电子病历、体检信息等数据，还涉及到公共卫生管理信息平台特别是疾控部门的医疗信息，这些医疗数据对防控区域性爆发的流行病、疾病间关系的发现具有很大的意义。随着大数据技术的发展，对医疗数据的分析及挖掘越来越被重视。

心血管疾病是一种严重威胁人类特别是中老年人健康的常见慢性病，发病时不易察觉、极易危及生命、易导致多种并发症、疗程长且难以治愈等特点，即使应用目前最先进、完善的治疗手段，仍可有50%以上的心脑血管意外幸存者生活不能完全自理，全世界每年死于心脑血管疾病的人数高达1500万人，居各种死因首位，如图1-1是近年来农村、城市各种疾病发病所占的百分比，可见心血管疾病是严重威胁人类身体健康的最主要因素，所以做到早预测，早治疗，提高心血管患者的生存率显得极其重要。



图1-1 心血管疾病在农村、城市的发病率

通过疾病预测评估过早干预用户可能患有的疾病，提高患者生存率一直是医学领域的研究热点，以前人们利用人口统计学、医疗条件、生活常规等基本信息来计算发展某种疾病的可能性，这种计算是使用基于方程的数学方法和工具完成的，一般涉及到的很少的变量和数据[32]。随着机器学习、大数据技术的快速崛起，不仅能够处理大量医学数据而且可处理大规模的变量，充分挖掘了医学数据潜在的规律，提高疾病预测的准确性。

监督学习作为机器学习领域中最多研究、应用最广泛的方法之一，通过己有的训练数据样本学习一个模型，使该模型对任意输入的待预测样本，都能得到一个好的预测输出结果。传统的单标签分类算法是把研究对象当作具有明确、单一的语义，对象被标注唯一的类别标签,已经取得了巨大的研究成果。由于现实世界的很多对象往往具有多义性和模糊性，并不具有唯一的语义，例如一张风景图片包含了海洋、沙滩、人三大类标签，一位病人可能同时患有心衰、心梗、脑卒中等多种心血管疾病，此时分类算法必须能够准确的识别该对象中的多个标签，而单标签算法无法准确描述一个对象可能和多个标签相关的问题，因此多标签分类（或多标签学习）研究应运而生。在多标签学习中每个对象由一个样本描述，该样本具有多个而不是唯一的类别标记，学习的目标是将所有合适的类别标记赋予未知样本。多标签学习广泛应用于图像标注、生物医学、文本分类等众多领域，不同的领域对多标签学习有不同的要求。

医学领域中真实医疗数据保密性强而不容易获取，同时这些数据内部存在极其复杂的关系且具有重大挖掘价值，为现代医疗发展起着必不可少的作用，因此医学领域的数据分析和研究一直是社会关注的焦点和相关学科的重点研究对象。近年来随着大数据技术的崛起，数据挖掘技术不断趋于成熟，在现代医疗数据研究领域得到了广泛应用。多标签问题最初提出的重要一部分原因就是医学领域数据挖掘的迫切需求。由于各种心血管疾病间复杂的医学关系、发病的先后关系及伴随的多种并发症等难以分析的特性，多标签分类算法不仅能预测单个疾病的发展情况，而且考虑了疾病间的复杂关系，能够识别患者可能患有多种疾病的风险。生物医学领域多标签分类问题普遍存在，心血管病因其时间长、隐匿性强、难以完全治愈等问题，在早期不易察觉常被患者忽视，多种并发症导致患者病情越发严重，以至于影响后期治疗。因此在医学基础上利用多标签分类算法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。由于疾病间的复杂性，多标签分类技术在医学领域的应用依然存在很大挑战。

## 国内外研究现状

### 机器学习在生物医学领域的发展现状

随着医疗大数据的出现，在计算机辅助医疗诊断技术的发展中，利用机器学习、数据挖掘方法与疾病相关的特征值来预测疾病变得越来越重要。例如Pena-Reyes 和 Sipper 等人利用模糊遗传算法预测乳腺癌，计算出超过 96%的准确率，Wang 等人利用人工神经网络模型和基于多层感知器来辨别口腔癌和口腔黏膜纤维瘤，得到了非常好的效果，Al-Ammar Barnes 利用有监督的聚类算法来预测癌症。[2]S Wan和MW Mak等人利用多标签算法解决蛋白质亚细胞水平上预测叶绿体蛋白的定位问题。[3]D Zufferey等人在已有的临床慢病数据上使用比较多种多标签分类算法来解决慢病预测问题。[4]W Zhang ,F Liu提出以ML-KNN为基础的新多标签分类算法解决药物副作用预测，取得了有效成果。医疗数据包括纯数据、信号、图像、文字等多种模式,其属性类型包括分类型、数值型或二者混合,数据中可能还包含了大量无用信息，因此,对医疗数据的降噪、剃选等预处理过程会比较复杂,挖掘过程交互性强,且可能需要反复多次。

总之,在医疗领域,多种数据挖掘算法都有了很好的应用,针对特定疾病问题,选择合适的挖掘方法,才能真正挖掘出符合临床实际的、有价值的知识。

### 多标签学习的研究现状

近年来，随着大数据技术的不断发展，机器学习得到空前关注及应用，为社会各领域做出来很大贡献。多年来经过学者们不断地深入研究，多标签分类问题有了许多显著的解决方案，并得到了很好的应用。根据文献[5]，总体上来说，多标签分类算法是单标签分类算法的扩展，主要分为PT（problem transformation）和AA（algorithm adaption）。

1. PT方法通过将多标签问题转换为一个个单标签分类问题进行处理，常见的有[6]BR(binary relevance）Binary、[7]Classifier Chains (CC)、 [8]Label Powerset ( LP)、[9]Hierarchy Of Multi-label learners(HOMER)、 [10]Random k-labelsets ( RAkEL)。BR算法为每一个标签训练一个的二元分类器，测试时，依次使用每个二元分类器判断测试对象是否属于对应标签。该算法简单直接，但是未考虑标签之间的相关关系。LabelPowerset(LP)将标签集合中的每个标签子集进行了二进制编码，转换为了单标签多分类问题，考虑了标签间的相关性，但是随着标签集合规模的不断扩大，标签编码将以指数形式增长，算法的复杂度变大。RAkLE（random k-labelset）算法解决LP算法中标签集数量过多的问题，该算法对标签随机分组，以组为限统计标签集，从而大大减少了标签集的数量
2. AA方法则是扩展已有的单标签分类算法使其能够处理多标签问题。基于单标签分类算法[11]AdaBoost.MI,Schapire等人提出了解决多标签分类问题的[12]AdaBoost.MH算法，该算法使用每个多标签训练数据生成ｑ(标签数量)个新的单标签训练数据，该算法的主要缺点是增加了训练数据的数量，加重了训练开销。[13]MLkNN通过改进knn算法，通过统计方法得出每个标签的先验概率，当输入一个未分类数据，对标签集合中的每个标签，分别计算该未分类数据具有该标签的概率，来预测该样本是否属于该标签。此外还有改进[14]C4.5算法的多标签决策树，基于支持向量机[15]（support vector machines, SVMs）和神经网络的改进算法等等。

多标签学习识别一种事物具有的多种特性更加切合现实世界，因此在图像、生物、文本等多领域得到了重视，例如Zincir-Heywood等人进行蛋白质功能分类，Li & Ogihara利用多标签算法分类音乐类别，Boutell 则用于情感语义识别[1]，多标签学习已经成为机器学习中重要分支。

## 论文主要研究内容

本文基于各医院、诊所等医疗机构的门诊病历数据，采用多标签分类算法建立心血管疾病预测模型，为获得可靠且符合实际意义的预测效果，主要研究工作包含以下几个方面：

1. 认识了解心血管疾病，以及国内外基于机器学习、多标签分类算法进行疾病预测的方法和案例。
2. 研究统计分析实验数据的特点，对心血管疾病预测模型进行研究，详细分析数据加载、确定并抽取研究对象、数据预处理、特征选择、模型建立、模型评价的处理方法。
3. 基于欠采样和过采样合并的思想，提出了ML-DARS(多标签双重自适应随机采样)算法，均衡了数据集分布，在此基础上采用混合策略应用多标签分类算法建立了心血管疾病预测模型。
4. 设计实现了基于心血管疾病预测模型的应用，将系统划分为数据采集模块、数据预处理模块、特征选择模块以及心血管病预测模块。内容如下：
5. 数据采集：原始数据由于数据量大且极不规则等特性，首先需要分析确定心血管疾病人群，然后提取例如脑卒中、心衰等属于终极心血管病的数据，再通过统计这些数据以及分布，确定要研究的对象，即用于多标签分类的各个目标标签。
6. 预处理：根据医学规则，本文的各研究对象（各目标标签）人群作为实验组，相应的其他标签人群作为对照组，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、属性转换等预处理操作。
7. 特征选择：为了预测结果的准确性、可靠性，对研究对象的特征进行降维处理，通过统计学方法和分类算法，进行二次特征选择，提取影响心血管疾病发展的有效因子，最终合并各标签的有效因子，保留效果最好的特征集合。
8. 心血管疾病预测模型：由于医疗数据分布的稀疏性、不平衡性，首先利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，提出了ML-DARS(多标签双重自适应随机采样)算法调整数据分布，直到满足多标签特性并得到更均衡的数据集。通过现有的多标签重采样算法和ML-DARS算法对比，将多个多标签算法应用于采样过的数据集上，通过多标签分类评价指标，分析不同的重采样算法导致的多标签分类效果的差异性。此外，为了提高分类效果，本文对现有的多标签分类算法进行了集成，针对心血管数据集特性提出了基于混合策略的多标签分类框架来构建心血管疾病预测模型，从样本近邻的局部角度和标签间相关性的全局角度来达到充分挖掘心血管疾病数据的目的。

## 论文的组织结构

本文主要工作分为四大部分：一是对心血管疾病的调研，多标签分类问题及解决方法的探讨与分析；二是本文介绍了用于研究的心血管疾病数据从提取开始、利用医学知识和数据预处理技术清洗数据、到特征选择以及最终应用于多标签分类算法等过程；三是深入理解多标签分类中的类别不均衡性问题，并在此基础上进行调研，通过学习已有的解决多标签不均衡性问题的方法，本文提出了一种多标签双重自适应性采样算法ML-DARS；四是利用现有的多标签分类算法构建基于混合策略的心血管疾病预测模型。具体结构如下：

第一章 主要介绍心血管疾病研究的重要性，多标签学习的研究现状，以及机器学习技术在医学领域的发展情况。

第二章 介绍多标签学习的定义，多种模型评价指标，现有且常用的多标签分类算法，在此基础上详细探讨了数据不均衡性问题及多标签领域不均衡问题的研究现状，同时描述了现有的评价多标签数据集分布的指标以及不均衡度。

第三章 考虑到医学数据复杂性以及专业性强等特点，本章对真实的医学首先做了深入的分析统计，确定了心血管疾病的多标签数据集，然后进行了一系列复杂的预处理过程和特征选择过程，为应用于多标签分类算法等全程的预测模型建立奠定了扎实的基础。

第四章 首先针对多标签不均衡性问题做了深入探讨，详细介绍了现有的解决多标签不均衡性的采样方法，为加快采样过程且获取更加均衡的数据集，提出了一种将欠采样和过采样相结合的多标签双重自适应随机采样算法，将其应用于公共多标签数据集和心血管疾病数据集，并在此基础上与现有的采样算法进行实验对比分析。

第五章 分析了心血管疾病数据集的自身特性，提出了基于混合策略的多标签分类框架构建心血管疾病预测模型，并对实验结果进行了分析。

第六章 对本文工作进行总结与评价，指出了研究中存在的不足，以及下一步研究方向。

# 多标签分类相关研究

## 单标签分类的局限性

机器学习中分类问题是典型的监督学习，通常分为两大步骤，一是训练模型，二是模型预测。如图2-1所示，训练模型时，首先将训练样本表示成模型可识别的特征向量，然后对特征属性进行特征选择，获得影响分类类别的有效特征，利用分类算法获得一定的泛化误差内尽可能拟合训练数据的分类模型，然后将测试样本表示成与训练样本相同的特征向量，通过分类模型，计算出测试样本相关的类标签。

测试样本

训练样本

分类模型

特征表示

特征选择

预测结果

图2-1 分类过程

传统的单标签分类算法将一个样本划分到唯一、特定的一个类别中，例如预测天气是否下雨、判断一个人的职业是教师还是警察等。然而现实世界中，我们通常遇到判断预测一种事物的多种语义，像一幅图片中同时包括海洋、蓝天、白云、树等等，一首歌曲通常被归到快乐、情感等多种类别当中，多标签学习面向于给定一个样本，将该样本通过分类算法，归类到相关的类标签中，可能是一个也可能是多个。

类别不均衡性指训练过程中某一类别的样本数目远超过其他类别的样本数目，导致分类模型效果变差甚至无效。现有研究中有了一些成熟的处理单标签分类中类别不均衡性方法，而多标签分类中考虑到标签间关系的复杂性，多标签类别不均衡性较单标签分类更为严重，处理该问题在多标签学习领域越来越被重视。

## 多标签学习方法

在单标签学习中，一个样本仅属于一个类，而多标签中一个样本可能属于多个类别，多标签问题可看成单标签学习问题扩展而得到的更加广义和复杂的分类问题。接下来介绍多标签学习的定义、相关的评估方法和已有的分类算法。

### 多标签学习的形式化描述

为了形式化描述多标签学习问题，设表示d维实例特征空间，为该空间的有限标签集合，即有M种可能的标签，多标签训练样本集 ，表示d维特征向量， 表示对应的第i个样本所关联的标签集合。多标签学习的目标是根据训练得到的函数，将测试样本集中的每一个样本, 获得与其相关的标签集合。



在多标签学习问题中，使用分类算法前，首先要衡量该数据集是否具有多标签分类的必要，因为它最终影响多标签分类效果，例如该数据集中的标签向量大多数仅有一个标签有效，收到的多标签分类效果非常微弱。该为了描述多标签数据集的特征，几种非常有用的多标签数据集衡量指标，标签基数见公式（2-1），表示每个样本平均的有效标签个数，一般来讲该值约大于等于2，具有较好的多标签数据集特性，对应的标签密度如公式（2-2）所示。另一种常用的多标签测量方法为标签差异性，衡量样本空间中出现的不同标签组合的数量，公式见（2-3）。

（2-1）



（2-2）



（2-3）



### 多标签学习的统计评价方法

多标签分类效果评估方法不同于单标签分类，学术界目前已经有许多成熟的多标签分类评估方法，大体上分为两大类，一种是二元分类评估，另一种是基于标签相关性排序的分类评估[15]。为了定义评估方法，给一样本，标签预测集合设，预测单个标签对应的排序函数为。与样本最相关的标签排序越靠前rank(1)，最不相关的标签排序越靠后rank(M)。



1. 二元分类评估

许多基于二元分类评估方法在于计算样本集的真实标签与预测标签的差异性的平均值,常见的Hamming loss[16]、Subset Accuracy[17]。另一种是对单个标签进行评估，然后取所有标签评估结果的均值，例如micro-averaged 、macro-averaged。在这里可又分为基于样本的评估和基于标签的评估方法。

1. 基于样本的评估方法

Hamming loss的定义如公式(2-4)：

（2-4）



∆表示两个标签集对应标签的差异性，即逻辑上的异或操作，代表相关标签被预测为不相关标签的比例，一定程度上值越小分类效果越好。

Subset Accuracy的计算公式(2-5)：

（2-5）



上式中，，该式严格要求预测标签集精确匹配真实标签集， 即预测标签等于真实标签的样本数所占比例，该值越大越好。



Precision, Recall, F1, Accuracy的定义见公式(2-6)：

（2-6）



1. 基于标签的评估方法

Micro-averaged和macro-averaged很类似，都是基于在单个标签上的二元评价方法。设二元评估方法为，其中代表样本集在单个标签上预测的真正类，真负类，假正类，假负类。Micro-averaged和macro-averaged的计算方法如公式(2-7)所示：



（2-7）



这些指标越大越好，最优值是1。

1. 排序评估

One-error评估方法表示排序最靠前的标签不属于样本的相关标签集合中的一员，对应的样本所占比例如公式(2-8)所示：

（2-8）



其中



Coverage指标见公式(2-9),指首先取样本的所有相关标签排序的最大深度，然后在所有样本上取平均，得到样本集的平均深度。该值越小说明分类时相关标签排序越靠前，分类效果越好。

（2-9）



Ranking loss的公式定义如(2-10)：

（2-10）



上式中 是 yi 的补集,即样本 xi 的不相关标签集合。对于单个样本而言,其排序损失值就是它的所有相关标签与不相关标签对中,发生排序错误的百分比。因此 Ranking Loss 就是所有样本的平均排序损失值，该值越小越好。



### 多标签学习方法

随着多标签学习技术的不断发展，许多重要的多标签分类算法被提出。这些算法总体上被分为两大类：基于问题转换的方法（Problem transformation methods）和算法适应性方法（Algorithm adaptation methods）。问题转换方法将多标签问题转换成已有的、成熟的单标签学习问题，典型的算法有Binary Relevance(BR)、classifier chains(CC)、Label Powerset(LP)、Random k-labelsets(RAKEL)、Hierarchy Of Multi-label learners (HOMER)。算法适应性方法则是对已有单标签分类算法进行修改来解决多标签分类问题，例如Multi-Label k-Nearest Neighbor(ML-KNN)、AdaBoostMH。下面进行详细探讨。

1. 基于问题转换的方法
2. BR

该算法非常流行的转换方法，对中出现每个不同的标签，学习到M个二分类器，该二分类器通常从单标签分类器中选取，例如决策树、SVM、朴素贝叶斯等。将原始数据集划分为M个包含所有原始样本的数据集，若原始样本中包含标签，中对应该样本标记为正类，否则为负类。当新样本到来时，BR输出被M分类器预测输出的类标签的组合。表2-1代表原始数据集，转换成BR算法要求的数据集如表2-2，将表2-1数据集划分为四个不同的二分类数据集。



表2-1 原始多标签数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **样本** | **属性** | **标签1** | **标签2** | **标签3** | **标签4** |
| 1 | X1 | 1 | 0 | 0 | 1 |
| 2 | X2 | 0 | 0 | 1 | 1 |
| 3 | X3 | 1 | 0 | 0 | 0 |
| 4 | X4 | 0 | 1 | 1 | 1 |

表2-2 BR转化后的数据集

|  |  |  |
| --- | --- | --- |
| **样本** | **属性** | **标签1** |
| 1 | X1 | 1 |
| 2 | X2 | 0 |
| 3 | X3 | 1 |
| 4 | X4 | 0 |

|  |  |  |
| --- | --- | --- |
| **样本** | **属性** | **标签2** |
| 1 | X1 | 0 |
| 2 | X2 | 0 |
| 3 | X3 | 0 |
| 4 | X4 | 1 |

|  |  |  |
| --- | --- | --- |
| **样本** | **属性** | **标签3** |
| 1 | X1 | 0 |
| 2 | X2 | 1 |
| 3 | X3 | 0 |
| 4 | X4 | 1 |

|  |  |  |
| --- | --- | --- |
| **样本** | **属性** | **标签4** |
| 1 | X1 | 1 |
| 2 | X2 | 1 |
| 3 | X3 | 0 |
| 4 | X4 | 1 |

BR算法每个标签独立的建立二分类器，可并行进行，具有简单高效性。然而该算法基于标签间相互独立的前提，忽略了标签间的相关性，丢失了许多的重要信息。同时由于标签集中有多个标签，导致转换后数据集正样本数目严重小于负样本，正负样本比例严重失衡，出现数据不均衡现象。

1. LP

LP充分利用了标签间的相关性，考虑了多标签中每组可能出现的唯一标签，对标签集进行二进制编码生成新类别，然后用单标签分类中多分类任务来完成。如表2-3， LP中新类别的生成。

表2-3 LP转换后的新类别

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **样本** | **属性** | **标签1** | **标签2** | **标签3** | **标签4** | **新类别** |
| 1 | X1 | 1 | 0 | 0 | 1 | 9 |
| 2 | X2 | 0 | 0 | 1 | 1 | 3 |
| 3 | X3 | 1 | 0 | 0 | 0 | 8 |
| 4 | X4 | 0 | 1 | 1 | 1 | 7 |

对于M个标签集合，理论上最多可生成2q 种不同类别，随着标签规模的不断增大，LP问题生成新类别的规模呈指数级别增长，不仅增加分类的难度，生成新标签时，有的标签对应的样本数目很大，有的的样本数目非常小，导致严重不均衡现象，致使预测效果变差。同时该方法无法学习到未出现的新类别，上表中将1,2,3样本作为训练集，4样本作为测试集，其转换后的新类别7未出现在训练集中，分类器不可能将其正确分类。

1. RAKEL

考虑到LP算法在大规模训练样本和标签上的计算复杂性及预测性能问题，RAKEL算法对LP算法进行了改进。RAKEL将原始标签集随机划分为一些小的标签子集，每个标签子集包含k个标签，且标签子集之间可能会有重叠现象，为每个标签子集使用LP方法将多标签分类问题转换为单标签中的多分类问题。若RAKEL选择m个分类器， 对一个新样本，可得到m个预测标签组合，预测组合公式如(2-11)。

(2-11)



表示标签在k个标签子集中实际出现的频率，表示标签在所有分类器中可能出现的最大频率，当标签出现的频率超过它所能获得的最大频率的一半时，就认为该标签是相关的。

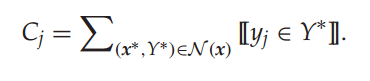


1. 算法适应性方法中的ML-KNN算法

ML-KNN的基本思想是采用k近邻技术处理多标签学习问题，利用最大化后验概率规则(MAP) 推理出待预测样本的标签信息。一未知样本，记表示样本集D中k近邻样本集。一般来讲，样本间相似度使用欧氏距离来度量。对于标签j, ML-KNN计算样本的标签出现在近邻中的次数，公式如(2-12)：



(2-12)



表示样本中出现标签事件，表示样本的近邻中存在个标签 时，出现的后验概率，相对的未出现的后验概率为，根据MAP规则， 预测标签集合由是否大于决定，如公式（2-13）所示：



（2-13）



根据贝叶斯理论，可得公式(2-14)：

（2-14）



表示出现的先验概率，可通过计算每个标签出现在训练样本集的频次来估计。后验概率利用似然估计计算，标签出现在训练样本集中的次数和k个邻居中有个标签来决定。类似，即标签没有出现在训练样本集中的次数和k个邻居中有个标签来决定。



ML-KNN继承了懒惰学习和朴素贝叶斯的优点，即决策边界可以自适应地调整，由于每个类标号是基于先验概率进行估计的，所以该算法对类别不均衡性程度不敏感，不足之处在于为每个待预测样本计算到全局已知样本的距离，计算量大。

## 标签不均衡性处理

### 单标签分类中类别不均衡性问题

在现实中有很多类别不均衡问题，它是常见的，并且也是合理的，符合人们期望的。如在欺诈交易识别中，属于欺诈交易的应该是很少部分，即绝大部分交易是正常的，只有极少部分的交易属于欺诈交易。又如，在客户流失的数据集中，绝大部分的客户是会继续享受其服务，只有极少数部分的客户不会再继续享受其服务。在监督学习领域，因为分配给每个类标签样本数量存在差异，我们经常遇到数据集不均衡现象，为降低全局误分率，分类器偏向于大类样本，损失了小类样本，分类器会大大地因为数据不平衡性而无法满足分类要求，因此在构建分类模型之前，需要对分类不均衡性问题进行处理。传统的单标签分类中对不均衡性问题进行了深入研究，如数据重采样，即在数据预处理阶段，通过删除部分大类样本或者增加小类样本来均衡数据集的分布。重采样技术独立于特定的分类算法，实践证明了其有效性。此外还有算法适应性，代价敏感性分类，它们则依赖于特定的分类算法。算法适应性方法是通过修正现有的分类算法处理数据不均衡问题，代价敏感性分类则结合算法适应性，在分类过程中采用代价敏感策略，误分类小类别样本的代价要大于误分类大类别样本的代价，算法更倾向于小类别样本。

### 多标签类别不均衡性探讨

单标签分类问题中一个样本仅属于一个标签，而多标签分类中大多数样本同时属于多个标签，这些标签数目通常在几十个到几百个之间，因此多标签学习中的类别不均衡性更为严重。尽管许多多标签数据集中出现大规模的标签集合，但是每个样本通常仅属于该标签集合的一小部分标签。图2-2展示了常见的7种多标签公共数据集中每个标签的样本比例，很容易看出大多数数据集中存在2到4个出现频率很高的标签，剩余的标签对应的样本数目则比较少，也就是说大多数标签被少于5%的样本所表示。

目前度量多标签数据集不均衡性的常用指标有两种方法，IRLbl度量数据集中每个标签不均衡度，定义如公式（2-15），利用样本集中出现最多的类标签的样本数目与当前标签的样本数目之比来表示该标签的不均衡率。MeanIR是对所有标签IRLbl值的加权平均，公式如（2-16）

， （2-15）



（2-16）



一般说，meanIR越大意味着数据集的不均衡程度越大，IRLbl(l)<meanIR认为标签l为大类标签，IRLbl(l)>meanIR则认为l属于小类标签。多标签数据集中存在多个大类别，也存在多个小类别，所以多标签数据集和传统数据集的不均衡性具有本质区别。多标签数据集中，经常出现小类标签和大类标签共同出现在同一样本中，这增加了多标签不均衡问题处理难度。

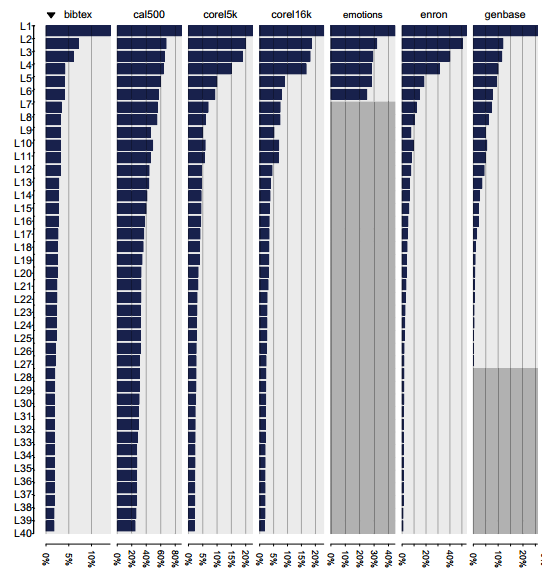


图2-2 已有多标签数据集统计

### 多标签类别不均衡性的问题处理方法

跟解决单标签分类中不均衡性问题类似，多标签类别不均衡性方法也主要有三大研究方向：算法适应性、集成方法、重采样技术。下面分别介绍：

1. 算法适应性

该方法经常基于多标签分类算法为每个标签赋予不同权重，缓解不均衡性问题。

如在文献[17]中人类蛋白质的定位预测问题面临高度不平衡的细粒度问题，解决方法是基于非参数概率模型，结合协方差矩阵获得标签相关性和每个标签相关联的加权系数来修复不平衡的问题。MIMLRB[18]F算法是基于径向基神经网络（RBFN）的多实例多标签细粒度的分类算法。MIMLRBF对RBFN进行优化，来解决多标签分类问题及多标签数据集的不平衡问题。MIMLRBF中通过调整连接隐层与输出层的权重来适应每个标签的单个偏置。Min–Max-Modular network[19]算法将分类任务划分成几个更小的任务，不同策略确保更小任务面临的数据不均衡性低于原始任务，这些子任务使用SVM算法处理。算法适应性方法依赖于特定算法，而且它们应用于特定领域。

1. 集成方法

基于分类器集成技术在多标签学习领域已经展示了它的优势，例如RAKEL、ECC、HOMER算法都获取了很好的性能，同样该方法常常用于解决不均衡问题。文献[20]利用RAKEL、ECC、MLKNN等算法为基础构建一个混合集成框架，缓解不均衡问题。对于传统分类问题中不均衡解决方法，集成BR分类器的方法解决多标签中不均衡问题，该算法名为BR-IRUS[21]， 思想是在每个标签上训练多个分类器，每训练一次使用所有小类样本和部分大类样本。这些方法的主要弱点是大量分类器的训练降低效率。

1. 重采样技术

重采样技术独立于多标签分类算法，基于数据预处理技术创建更为均衡的数据集，是最为广泛使用的数据均衡方法，现有的多标签重采样技术几乎都是受传统单标签学习中重采样方法影响，即欠采样、过采样、小样本合成技术。

目前多标签学习领域处理多标签不均衡性问题的采样方法被提出，最早的是欠采样LP-RUS和过采样LP-ROS算法，基于LP转换多标签数据集的方法，将LP转换后的每个标签集作为一个类别定义，LP-RUS随机删除大类别样本，LP-ROS随机复制小类样本增加小类样本的数目，该算法简单有效，但由于该算法中最大可生成种不同新类别，增加了算法复杂度。此外该算法不能解决这类极端问题：一个样本集中有502个样本，对应了502种不同的标签集合，也就是说样本集中的所有标签集都是不同的，虽然MeanIR指标显示数据不均衡很高，LP不能处理该问题，因为生成的新类别数目是相同的。为了解决LP-RUS和LP-ROS算法存在的问题，ML-RUS和ML-ROS算法对其进行了修改，单独评估原始数据集中每个标签的不均衡度，而非LP中生成的新类别，降低了复杂度。鉴于ML-ROS算法直接复制小类标签的样本方法虽然在一定程度上均衡了数据分布，但是也造成了大量重复样本导致的过拟合问题。为解决该问题，MLSMOTE算法利用传统单标签分类中小样本合成技术SMOTE，对训练集中小类样本通过插值产生额外的新样本，丰富小类样本同时避免过拟合问题。



## 本章小结

本章首先介绍了多标签学习的定义、评价指标以及相关方法。之后就类别不均衡问题进行了探讨，先描述了类别不均衡性问题，之后对传统单标签分类中的类别不均衡性现象及处理方法做了介绍，最后引出了多标签情况下的类别不均衡性探讨，解释了多标签类别不均衡的情况，类别不均衡性评价指标，和已有的多标签类别不均衡性处理方法。通过本章对多标签分类中一些知识点的介绍，为之后多标签学习应用于心血管疾病预测奠定了基础。

# 心血管病预测模型研究

## 模型建立的目标和步骤

本论文以医疗大数据项目为背景，利用大数据和机器学习技术，按照医学规则，从百万条高血压、糖尿病患者医疗数据记录的电子病例、疾控、体检数据中提取与心血管疾病相关的信息，并进行处理、分析，发现影响心血管疾病的潜在因子，在此基础上应用多标签分类算法对心血管疾病进行风险预测，以辅助医疗诊断。

医学数据的处理以其复杂性、不完整性、专业性强等多种特性，一直是各项研究中首要也是必不可少的重要环节。由于门诊数据和临床数据是经人工录入系统的，所以录入过程中难免出现一些失误或者意外情况，例如患者就医时对于某些医学参数不清楚导致该医学指标缺失，医生在录入数据过程中也可能出现失误，患者可以在多个医疗机构就诊而导致该患者的医疗数据产生冗余、不一致性。此外，这些原始的医学数据指标之间存在属性多样性，例如年龄为数值型属性、患者症状描述则为文本类属性。

由于支撑本文的项目中原始医疗数据包括了所有患者的数据，其数据量大，数据复杂，而并非仅是心血管疾病数据，为了使拿到的原始医学数据更好的应用于心血管疾病预测模型，本文先对原始数据进行分析统计，选取并确定特定的心血管疾病，提取所需数据并进行清洗预处理过程，处理完后还需要从大量的特征中选择有效的心血管疾病影响因子，最后才应用于预测建模并进行结果分析。其基于多标签分类算法的心血管疾病预测建模过程如图3-1所示。

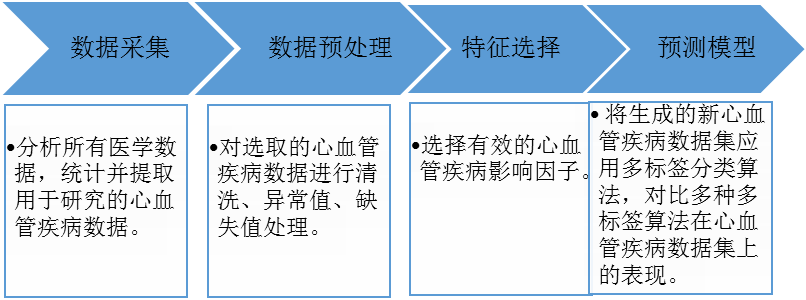


图3-1 模型建立步骤

## 心血管疾病数据采集

### 数据集分析

本文中的心血管疾病数据针对于高血压、糖尿病人群，是一个城市所有医疗机构所拥有的全部高血压、糖尿病患者的医疗数据，其中包括了所有患者个人信息、患病症状、各种健康指标等信息。表信息描述见表3-1，由于一个患者标识对应多条就诊记录，所以记录总数不小于患者数目。

表 3-1 数据来源说明

|  |  |  |  |
| --- | --- | --- | --- |
| 含义 | 记录总数 | 患者数目 | 相关数据 |
| 个人信息表 | 1204757 | 185712 | 性别、年龄、文化程度、婚姻状况、身高、体重、BMI等 |
| 高血压表 | 446883 | 123019 | 门诊机构、医院编码、高血压级别、呼吸频率、收缩压、舒张压、心率、各种并发症等 |
| 糖尿病表 | 120451 | 65798 | 呼吸频率、收缩压、舒张压、心率、各种并发症等 |
|  |  |  |  |

由于每位患者患病情况不同，所以高血压表、糖尿病表中存在大量病症信息，这些疾病种类繁多，处理起来十分麻烦，但是医学上对这些疾病都有统一的归类，鉴于本文定位于心血管病人群，所以这里仅对心血管疾病数据进行归类，我们首先根据医学规则进行ICD编码将所获得数据特征中属于同一大类的疾病进行合并，部分信息见表3-2，例如下表中肺心病这一类包括了医学上的原发性肺动脉高压、慢性肺源性心脏病等多种小类疾病组成。

表3-2 疾病归类

|  |  |  |
| --- | --- | --- |
| 疾病代号ICD | 疾病名称 | 所属类别 |
| I25\_000 | 被描述为动脉硬化性心血管病 | 稳定性冠心病 |
| I25\_100 | 动脉硬化性心脏病 |
| I25\_101 | 冠状动脉狭窄 |
| I25\_102 | 冠状动脉粥样硬化 |
| I27\_0 | 原发性肺动脉高压 | 肺心病 |
| I27\_8 | 特指肺源性心脏病 |
| I27\_9 | 慢性肺源性心脏病 |
| I11\_000 | 高血压心脏病伴心力衰竭 | 心力衰竭 |
| I13\_000 | 高血压性心脏病肾脏病伴心力衰竭 |
| I50\_000 | 充血性心力衰竭 |
| I50\_100 | 左心衰竭 |

### 研究对象及预测目标的选取

本文的研究定位是心血管疾病类人群，而上述疾病包括了心血管类、消化类、内分泌类等各种疾病人群，所以需要从中统计提取所需要的心血管类人群，经过对数据源的疾病合并归类后，得到的心血管疾病数据统计结果如表3-3，这些疾病的人群可重复，因为一位患者可能患有多种疾病。

根据上表的统计结果，选择心血管疾病作为研究目标，有的疾病对应的人数太少不考虑加入多标签集合，例如肺栓塞仅21人，不具备预测能力，考虑到严重的心血管疾病危及人们的生命，所以对预测严重心血管疾病具有重要的研究价值，最终选择用于多标签学习中的研究对象如表3-4所示。

表3-3 初步数据统计

|  |  |
| --- | --- |
| 疾病名称 | 患者人数 |
| 严重脑神经疾病、意外死亡及后遗症 | 60050 |
| 心衰人群 | 1079 |
| 心源性猝死 | 128 |
| 脑卒中 | 7809 |
| 急性心梗 | 1030 |
| 肺栓塞 | 21 |
| 肾衰竭 | 1168 |
| 心肌缺血 | 1137 |
| 心脏功能病变 | 3062 |
| 冠心病 | 16761 |
| 高血压 | 123019 |
| 糖尿病 | 65798 |

表3-4 研究对象说明

|  |  |
| --- | --- |
| 心血管病 | 患者数目 |
| 脑卒中 | 7809 |
| 心衰 | 1079 |
| 心梗 | 1030 |
| 肾衰 | 1168 |
| 心肌缺血 | 1137 |
| 心脏功能病变 | 3062 |
| 冠心病 | 16761 |
| 高血压 | 123019 |
| 糖尿病 | 65798 |

### 数据采集实现

1. 数据采集平台

原始数据由于数据量大且极不规则等特性，通过结构化组织存储在Oracle数据库中，包括个人信息表、高血压相关表、糖尿病相关表等，如表3-1所示，这些表存储的数据量都在几十万以上，不论是表关联还是数据统计，数据量更大，在计算量、计算速度上难度很大。

为此，利用现在成熟的大数据存储处理平台hdfs、spark技术进行处理。Hdfs技术作为分布式大数据存储架构，具有高可靠性。Spark是适用于大数据的高可靠性、高性能分布式并行计算框架，支持内存技术、多迭代批量处理、流计算和图计算等多种范式，Spark 内存计算框架适合迭代式算法和交互式数据分析，能够提升大数据处理的实时性和准确性。本次数据处理以hdfs为后台存储原始数据的地方，spark则作为计算平台，从hdfs中读取数据并进行相关计算，数据采集处理的平台架构如图3-2所示。

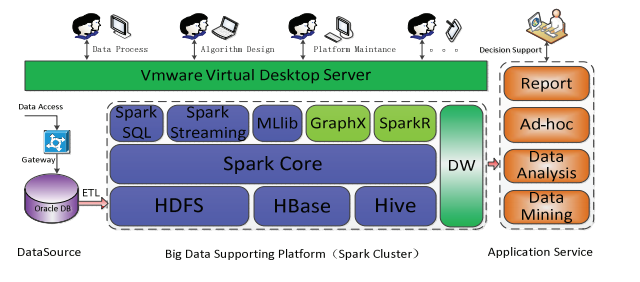


图3-2 数据采集处理平台

本次分布式处理平台中，spark集群配有13台机器，一个作为主节点，4核16GB内存，500GB磁盘，其他作为计算节点，8核12GB内存，500GB磁盘。

本次数据采集过程将原始数据从Oracle 导出并加载到hdfs中存储，利用医学知识、spark平台、Java语言，提取例如脑卒中、心衰等属于终极心血管病的数据，再通过统计这些数据以及分布，确定要研究的对象，即用于多标签分类的各个目标标签。

1. 数据采集过程

首先关联个人信息表、高血压表、糖尿病表，初步获得患者的所有患病信息，然后根据医学定义读取图3-3各文件的ICD码统计各心血管疾病数据量结果如图3-4，通过分析，决定了最终的目标标签如表3-4，并合并目标标签的各ICD码作为一大类疾病的结果值，认为只要子ICD中出现该病就认为患者患有该标签代表的心血管疾病。核心代码如图3-5，展示了关联表过程和提取ICD码对应的目标疾病人群。

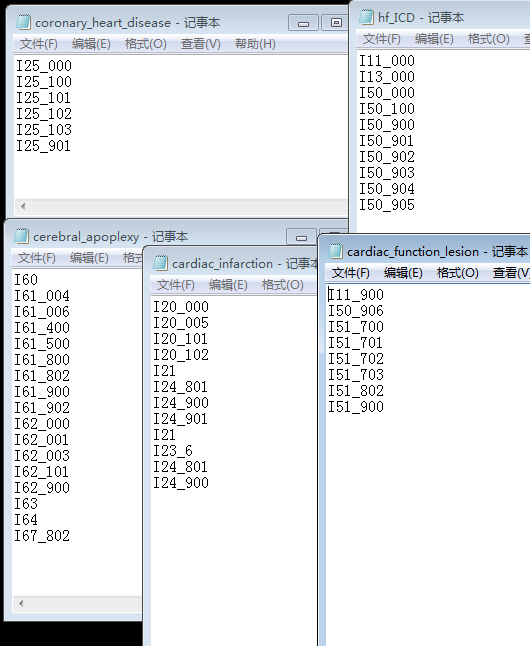


图3-3 各心血管疾病统计

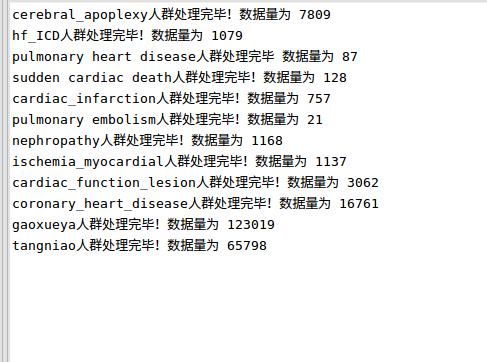


图3-4 统计各心血管疾病数据量

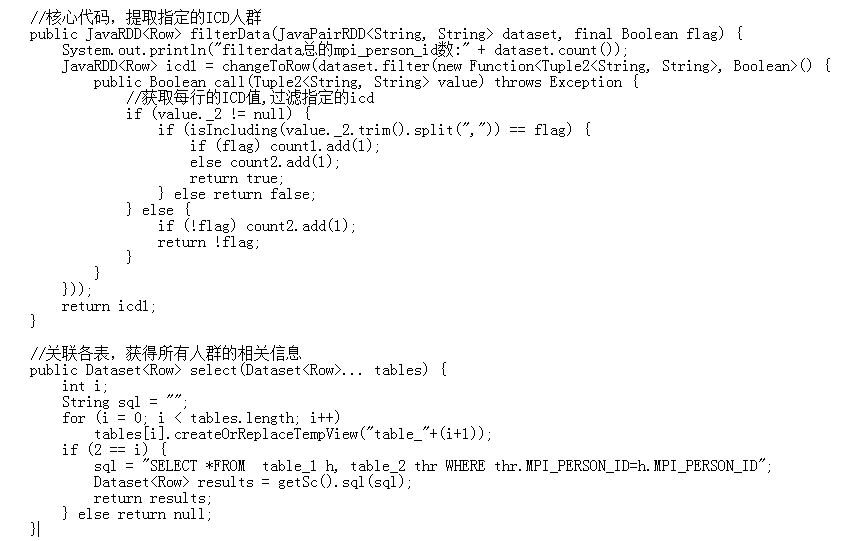


图3-5数据采集核心代码

## 患病记录数据预处理

### 数据清洗与有效数据提取

本文中所用到的真实数据中，患者患病记录的医学特征分为两大类，一是一般身体指标如年龄、性别、血糖等，这类特征数目比较少，对应的属性类型既有数值型也有文本型比较混杂，另一种则是患者对应的患病种类如一个患者可能患有心律失常、心肌缺血等多种并发症，这类特征每个疾病ICD编码单独作为一列，属于0-1属性，即患有该病为1，没有则为0，因为一位患者一般来说患的疾病数很少大约2-3种，而所有患者患病情况千差万别，所以这类特征很稀疏。

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，此外鉴于患者门诊数据中一位患者可对应多个门诊记录，不符合本文中的多标签数据集的要求即一位患者对应一条记录，所以医学数据处理起来比较复杂。很明显，这里主要处理的特征集中于上述提到的一般身体指标。

根据医学规则，首先根据数据集特性，删除对模型无意义特征例如机构编码、档案流水号、终止管理日期、随访流水号、并发症的取值全为0等。删除冗余性特征，仅保留一个，例如出生日期和年龄只需保留年龄即可，许多特征既有该指标编码又有该指标名称，只需保留指标编码就可以了。大量的缺失值使得机器学习的效果不理想，无法学习到好的模型，因此需要初步统计每一个特征的非缺失值数目，删掉缺失值在70%以上的特征，例如婚姻状况、血型等。

### 异常值处理

对医学特征中出现不符合常识的异常数据进行检测、处理，鉴于医学数据的严谨性、标准化特性，使用人工检测表方法规定生物学特征的取值范围，不在范围内的被视为异常值，为了尽可能保留研究对象的数据量，除了日期数据异常进行删除操作之外，其他特征异常值视为缺失值处理。人工检测表部分信息如下表3-5所示。

表3-5 检测表部分取值

|  |  |  |
| --- | --- | --- |
| 特征名称 | 正常取值范围 | 特征类型 |
| 性别 | 1男 2女 | 离散型 |
| 年龄 | 4到115之间 | 数值型 |
| SBP收缩压 | 80到200之间 | 数值型 |
| DBP舒张压 | 40到160之间 | 数值型 |
| 心率 | 10到220之间 | 数值型 |
| BMI（身高体重比） | 10到50之间 | 数值型 |

由于并发症日期(确诊日期)唯一但随访日期(复查日期)不唯一，所以一个患者标识对应多条记录，优先提取随访日期距离并发症日期最早且在并发症日期之前的记录，因为我们需要获取待预测疾病可能存在的先验条件，即每个患者对应一条记录。

### 缺失值处理

缺失值填充过程，首先，有的特征属性为文本属性例如症状编码记录患者身体情况，经分析发现主要集中于特定的几个词，所以该属性可看做离散属性，进行维度扩充，对于每一个值，出现该特征的记为1，否则记为0，这样就不存在缺失值的情况。由于一位患者对应多条随访记录，所以使用该患者相关的记录数据处理缺失值相对更具有可靠性，因此数值型属性以一位患者多条记录中的其他非缺失值的平均值填充，离散型属性则以一位患者多条记录中的其他非缺失值出现次数最多者进行填充。因为存在一位患者的多条记录都为空值的情况，所以对剩余的缺失值处理过程根据每个特征的特性分为两部分，一是不变性特征例如体重、身高、性别等均值或者众数填充，二是易变性特征例如心率、空腹血糖等零值填充。

由于各个特征间属性值因为实际意义不同存在较大差异，需要对数据进行归一化处理，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。数据归一化方法常有z-score归一化，将特征处理后，数据均值为0，标准差为1，其公式（3-1）所示，此外还有离差标准化，将原始数据线性变换后使结果落到[0,1]区间转换函数如公式(3-2)所示。鉴于离差标准化方法有一个缺陷就是当有新数据加入时，可能导致max和min的变化，需要重新定义，所以对此医疗数据采用Z-score方法进行数据归一化处理。

(3-1)



为所有样本数据的均值，为所有样本数据的标准差。



(3-2)



其中max为样本数据的最大值，min为样本数据的最小值。

### 数据预处理实现

本次数据预处理实现在3.2节基础上，依然利用spark计算平台来完成。该过程的ETL流程代码如图3-6所示，根据得到的心血管疾病数据集，为获取有效数据，首先删除明显的无效特征，然后根据人工检测表处理异常值，核心代码如图3-7所示，离散型和连续型特征选择零值填充，而异常日期数据无法预测进而删除该记录。最后考虑到缺失值需要用患者的多条记录进行处理，所以将两者一起处理，如图3-8，final\_data函数在取患者最早的随访记录时，考虑到该记录中存在的缺失值，所以同时又根据其他记录处理该缺失值，即调用了sum\_null函数见图3-9，average函数则是对连续属性均值处理，离散属性众数处理。



图3-6 数据采集预处理流程代码



图3-7 异常值检测处理

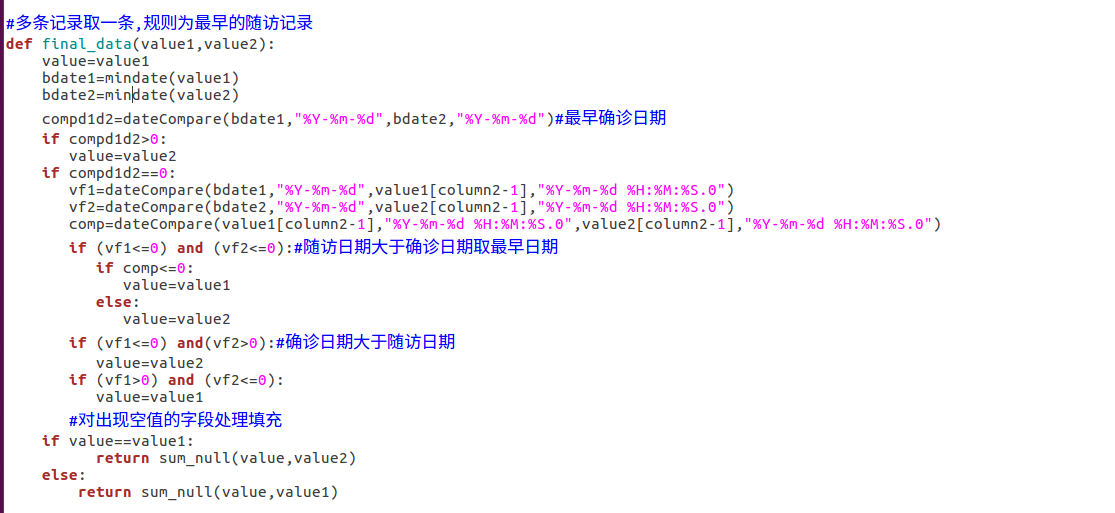


图3-8 缺失值处理



图3-9 多条随访记录数据记录取一条

## 心血管疾病数据特征选择策略

### 特征数据提取

在数据预处理阶段，去掉了一部分很明显像用户ID这样的无意义特征，但是依然存在某些对类别区分度不大的特征，在进行分类学习时，过多的特征可能会导致训练效果下降。为了预测结果的准确性、可靠性，需要对数据进行进一步降维处理，提取影响心血管疾病发展的有效因子。此外，考虑到医学领域数据的严谨性、保守性、复杂性，本文选择的心血管疾病间可能存在非常密切的相互作用，同时多标签学习领域的特征选择方法尚未成熟，所以本文采用统计检验和分类算法相结合，提出了基于多标签分类的心血管疾病数据特征二次选择策略。

### 特征数据的二次选择

该策略将已预处理好的基于心血管疾病的多标签数据集使用多标签学习领域常用的转化方法BR来对标签集合中的每个标签单独处理。处理过程分两步，如下：

1. 首先对研究对象进行特征相关性分析，去除共线性强的特征，选择与标签相关性较强的特征，常用的统计学方法有pearson相关系数、spearman秩相关系数，两种方法都是度量两个随机变量的相关程度，pearson相关系数用协方差除以两个变量的标准差得到的，介于-1到1之间，当两个变量线性关系增强时，当一个变量增大，另一个变量也增大时，表明它们之间是正相关的，相关系数大于0；如果一个变量增大，另一个变量却减小，表明它们之间是负相关的，相关系数小于0；如果相关系数等于0，表明它们之间不存在线性相关关系，该方法限于两变量呈线性相关关系，如果是曲线相关可能不准确，此外两变量须符合正态分布。Spearman相关系数对原始变量不做要求，适用范围广，通常被认为是排列后的变量间的pearson线性相关系数。本文考虑到心血管疾病特征的复杂性及特征存在的稀疏性，选用spearman作为检验手段，此过程选择出与每个疾病相关性强的特征，保留p<0.05的特征。
2. 经过特征相关性选择，获得了与特定单标签疾病相关的特征，去除了共线性特征，一方面为了验证第一次特征选择的有效性，将选出的有效特征作为特定标签的属性，应用逻辑斯特回归分类算法分别评估单个心血管病的分类效果，另一方面，由于第一次处理后每个标签的特征维度平均达到了90多个，所有标签特征组合起来特征维度依然很高，逻辑斯特回归是机器学习领域常用的有效分类算法，其模型公式如（3-3）。数据统一分布在0-1之间即Y∈(0,1),可以看出模型中权重ω越大，表示该特征对预测结果的贡献度越大，当|ω|越小，说明该特征无法对预测结果产生影响，从中剔除，所以这里用逻辑斯特回归模型产生的特征权重进行进一步特征选择来获得更为干净的有效因子集合。根据特征权重由高到低排序，选择排序靠前的特征作为该疾病的度量指标，权重越大对目标疾病的贡献程度越大。

(3-3)



根据第一次特征选择的结果，对所有疾病进行了二次特征选择，获得每个标签对应特征的权重，根据权重对特征排序，最后获得每个特征在所有目标标签上的平均排名，根据实际情况合并所有心血管疾病的特征。

### 特征选择实现

以心衰为例部分结果如表3-6，例如该表展示了心律失常的相关性，spearmanr\_pval值小于0.05，95%的置信度认为两个随机变量相关，表明心律失常与心衰间存在强相关性，而由于慢性缺血性脑血管病和心衰的spearmanr\_pval值大于0.05，无法证明两者间存在一定联系。

表3-6 心衰Spearman相关性分析

|  |  |  |
| --- | --- | --- |
| **特征** | **特征名** | **spearmanr\_pval** |
| **I67\_2** | 大脑动脉粥样硬化 | 2.21536E-12 |
| **DIZZINESS** | 头晕头痛 | 5.78954E-61 |
| **BIT16** | 代谢紊乱 | 9.32435E-58 |
| **BIT38** | 肺水肿 | 2.22117E-32 |
| **BIT33** | 肺炎 | 4.70524E-20 |
| **I49\_900** | 心律失常 | 3.36162E-29 |
| **BIT46** | 胸痛、呼吸异常 | 6.92032E-27 |
| **AGE\_NEW** | 年龄 | 2.99704E-18 |
| **BIT32** | 感冒引起的症状 | 0.0826336673 |
| **I67\_805** | 慢性缺血性脑血管病 | 0.1247875084 |
| **BIT40** | 多种相关的皮肤病 | 1.451383E-09 |

同样以心衰为例，部分结果如表3-7，年龄、心率等特征权重绝对值较大有利于预测结果，而皮肤病类、代谢综合征等权重过小对预测结果影响因素较小。

表3-7 逻辑回归特征选择

|  |  |  |
| --- | --- | --- |
| **特征** | **说明** | **模型权重** |
| AGE\_NEW | 年龄 | 0.06502474 |
| HEART\_RATE\_TIMES | 心率 | -0.05905261 |
| BMI | 体重身高比 | -0.09750807 |
| BIT33 | 肺炎 | 0.299 |
| DBP | 舒张压 | 0.10771338 |
| SBP | 收缩压 | 0.084 |
| I49\_900 | 心律失常 | 0.179 |
| BIT46 | 胸痛、呼吸异常 | 0.287 |
| BIT40 | 多种相关的皮肤病 | -0.00428 |
| G90 | 自主神经系统疾患 | 0.013 |
| E16\_803 | 代谢综合征 | 0.00508 |

最后，根据第二次特征选择结果，获得每个特征在所有目标标签上的平均排名部分结果见表3-8，均值是该特征排序的平均值，多个特征可能有多个相同的值，最后的综合排名是特征在所有标签的排名，最后保留了效果最好的296个特征集合。

表3-8 特征在每个标签的排名

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 目标标签 | I47\_1 | 年龄 | BMI | I48\_X02 | SBP | R45 | E16\_200 | E16\_803 | BIT16 | G20\_X00 |
| 脑卒中 | 12 | 14 | 26 | 23 | 56 | 153 | 67 | 143 | 74 | 19 |
| 心衰 | 33 | 6 | 75 | 77 | 58 | 7 | 26 | 61 | 62 | 52 |
| 心梗 | 3 | 8 | 54 | 77 | 32 | 121 | 16 | 14 | 26 | 39 |
| 肾衰 | 4 | 27 | 42 | 16 | 39 | 157 | 41 | 32 | 51 | 125 |
| 心肌缺血 | 36 | 13 | 9 | 76 | 27 | 17 | 28 | 12 | 47 | 14 |
| 心脏功能病变 | 10 | 70 | 27 | 24 | 31 | 8 | 103 | 26 | 55 | 126 |
| 冠心病 | 14 | 24 | 25 | 57 | 54 | 15 | 32 | 58 | 45 | 50 |
| 高血压 | 66 | 33 | 37 | 44 | 34 | 14 | 35 | 59 | 91 | 52 |
| 糖尿病 | 3 | 23 | 25 | 19 | 153 | 2 | 147 | 99 | 63 | 45 |
| 均值 | 20.11 | 24.22 | 35.56 | 45.89 | 53.78 | 54.89 | 55 | 56 | 57.11 | 58 |
| 综合排名 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

## 本章小结

本章对心血管疾病数据集应用于多标签分类算法之前对数据处理流程进行了详细说明，首先分析了数据集本身的特性、选取规则及确定研究对象，然后对数据集做了必要的数据清洗，最后为了更好的应用多标签分类算法，本章对数据特征进行了降维，根据心血管疾病数据集的特性，利用spearman统计和逻辑回归算法相结合进行了特征二次选择策略，获得有效的影响因子，此外还对各部分工作进行了实现。本章通过对心血管疾病数据集进行一系列的预处理操作，为后续工作打下了坚实的基础。

# 多标签双重自适应随机采样算法

## 引言

本文中心血管疾病预测模型是基于医院历史门诊数据中的诊断结果和患者自身的体检指标作为模型预测的特征集合。

由于一位病人一般并发症数目不超过3种，且一种疾病一般仅与少数其他疾病存在明显的相互关系，因而心血管疾病数据集中特征诊断因子存在稀疏性，数据分布存在不均衡等问题，使得分类算法偏向于大类样本，导致预测效果缺乏可靠性。

为此，本文在建立预测模型之前，需要解决多标签数据集的不均衡性问题。多标签分类问题中不仅涉及到多数类标签，而且还要考虑类标签间的相互关系，所以多标签分类比传统单标签分类要面临更为复杂的挑战。多标签分类过程中经常遇到类标签分布不均衡性现象，这些不均衡现象要比传统分类中的更为严重，加剧了多标签分类任务的难度。虽然多标签学习领域的不均衡性数据集到处可见，但是处理多标签不均衡性的方法近些年才被重视和研究。不论是传统的单标签分类还是多标签分类，重采样算法以独立于具体分类算法且成熟有效而得到了广泛研究和应用。

目前，解决多标签不均衡性问题的重采样方法是欠采样和过采样两种方法，这两种方法分开独立的处理数据不均衡性问题，实际上多标签数据集中大类样本和小类样本之间存在很大差距，如果仅对数据集进行欠采样，为了和小类样本达到均衡，大类样本数目变少，而小类样本过少很难学习到有效模型，导致最终的模型欠拟合。如果仅对数据集过采样，小类样本为了缩小和大类样本的差距，对小类样本进行复制或插值而产生的新样本数目超过了原始的小类样本数目，可能造成模型过拟合，有甚者产生的噪声数据对模型造成严重影响。

通过研究分析，本章提出了一种多标签双重自适应随机采样算法ML-DARS(Multi-label double adaptive random sampling algorithm)，并通过与其他已有的数据均衡性算法比较，证明了新算法的有效性。该算法考虑整个数据集的分布，从两个层面解决多标签不均衡性问题即欠采样过程和过采样过程相结合，降低数据集中大类样本的同时也适当的增加了小类样本的数目，防止了大样本信息大量丢失和小样本数据冗余问题。通过将该算法应用于心血管疾病数据集和公共数据集，ML-DARS算法产生的数据集符合实际的数据分布的特点，很好地解决了多标签数据不均衡性问题，提升了分类效果。接下来详细探讨已有的多标签重采样算法，在此基础上重点介绍ML-DARS算法，以及对应的实验结果和评价。

## 多标签重采样算法

目前多标签学习领域的不均衡问题解决方法集中于重采样技术，分为基于LP方法和基于ML方法，其中ML方法在LP方法的基础上进行了修正，更加有效的解决了多标签不均衡性问题。基于ML方法对标签集合中的每个标签评估不均衡性，如公式（2-12）、(2-13)，IRLbl>MeanIR, 将标签归到小类标签集合中，IRLbl<MeanIR则归入大类标签集合中，在此基础上采用不同策略对训练集重采样。基于ML方法具体包括ML-RUS欠采样算法、ML-ROS过采样算法、ML-SMOTE小样本合成算法，对此接下来将会详细介绍。

1. ML-RUS算法

ML-RUS算法主要思想是随机删除不在小类标签集合中的样本。包含小类标签的样本从要删除的样本集中排除出去，每次从删除样本集中随机选择一样本进行删除，该样本仅含有大类标签，之后检测大类标签样本集中每个标签的IRLbl是否大于初始的MeanIR，若是说明该标签对应的样本集变小了，将该标签从大类标签集合中删除，重复上述过程直到达到规定的阈值，完成欠采样过程。算法伪代码过程见算法1,该算法在一定程度上缓解了多标签数据不均衡问题。

算法1 ML-RUS

|  |
| --- |
| Input: 训练数据集D, 对应的标签集合L, 欠采样比例P  Output: 采样过后的数据集Y  (1) samplesToDelete=|D|\*P, P表示从训练集中删除的样本比例  (2) **for**  each label  **in**  L  **do**  (3) IRLbllabel =calculateIRLbl(D,label)  (4)  **If**  IRLbllabel < meanIR  **then**  (5) Label 加入maxBag 集合中  (6)  **else**  label 加入 minBag中  (7) **End for**  (8) **While**  samplesToDelete >0  **do**  (9) 从训练集D中随机选取一个样本x  (10)  **For** each minBagi  **in** minBag  **do**  (11)  **If**  xminBagi ==1 **then**  (12) Continue  (13)  **End for**  (14)  **For** each maxBagj  **in**  maxBag **do**  (15) 更新maxBagj的IRLbl  (16) **If**  IRLblmaxBagj >=meanIR  **then**  (17) 将maxBagj从maxBag中移除，并加入到minBag中  (18)  **End for**  (19) samplesToDelete—  (20) **End while** |

1. ML-ROS算法

类似于ML-RUS算法，ML\_ROS算法根据公式（2-12）、（2-13），首先将原始标签集划分为小类标签集和大类标签集，不同之处在于该算法对每个小类标签，找到小类标签对应的实例集合，遍历并在其中随机挑选一个实例进行复制，并将复制的实例添加到训练集中，当标签的IRLbl达到MeanIRL说明该标签对应的样本数目接近于平均样本样本数目，停止随机挑选的过程，并将该标签从小类标签集合中除去，重复以上过程直到满足指定的阈值，过采样过程完成。这种方法若直接复制样本的数目过多会造成模型的过拟合现象。

1. ML-SMOTE算法

由于ML-ROS算法直接复制样本而导致的过拟合现象，ML-SMOTE算法受单标签分类中小样本合成算法SMOTE的启发，根据出现在小类标签集合中的每个标签对应的实例集合来合成新的样本。这个新合成的实例以相关样本和对应的近邻为基础，形成了自己的特征集合以及对应的标签集合。因此ML-SMOTE主要为以下步骤：

1. 小类样本选取：该算法设定大多数多标签数据集至少有一个小类标签，小类标签选择与ML-RUS、ML-ROS一样，根据IRLbl和MeanIR的关系，找到对应的小类标签集合。
2. 获取近邻样本：首先获取小类标签对应的样本集合，从中任选一个样本作为种子样本，寻找距离该样本最近的k个实例样本。
3. 形成新样本特征集合：根据选择的k个近邻实例，新样本的特征值利用小样本合成算法来生成，从邻居中任选一实例，对于数值型属性，种子样本和选取的实例的特征值之间随机生成一个值，对于离散型属性，出现在所有邻居中次数最多的值作为特征值。
4. 标签集合的生成方式：新合成样本的标签集合生成方式可以是种子实例和近邻实例标签集合的交集或者是并集，也可以是邻居样本通过投票的方式超过半数的标签作为新样本的标签集合。

ML-SMOTE算法过程如算法2中的描述，基于邻居投票方式生成的标签集合效果最好。

算法2 ML-SMOTE

|  |
| --- |
| Input: 训练数据集D, 近邻数目k  Output：插值后的新数据集  (1) L 为训练集D对应的原始标签集合  (2) **For** each label  **in** L  **do**  (3) IRLbllabel = calculateIRLbl(D, label)  (4)  **If**  IRLbllabel > meanIR **then**  (5) label 加入小类标签集合minBag  (6) **End for**  (7) **For** each minBagi **in**  minBag  **do**  (8) minBagSamples = 获取minBag对应的所有样本  (9)  **For**  each sample  **in** minBagSamples **do**  (10) 计算sample 在minBagSamples 中距其他样本的距离记为 distances  (11) 根据distances对样本由小到大排序  (12) 获取最靠前的k个实例 得 neighbors  (13) neighbors中随机选择一个实例refNeigh  (14) synthSmpl = newSample(sample, refNeigh, neighbors)  (15) synthSmpl加入训练集D中  (16) **End for**  (17) **End for**  (18) **Function**  newSample(sample, refNeigh, neighbors)  (19) synthSmpl =new Sample  (20) **For** each feature **in**  synthSmpl **do**  (21)  **If** feature is numeric **then** //数值型属性插值处理  (22) Diff = refNeigh.feature – sample.feature  (23) Offset = diff \* randvalue(0,1)  (24) Value =sample.feature + offset  (25)  **Else** value =mostFreVal(neighbors.feature)//频次最多  (26) synthSmpl.feature = value  (27) **End for**  (28) Lblcounts = count(sample.labels, neighbors.labels) //计算种子样本和邻居样本中标签频次  (29)  **If**  lblcounts>(k+1)/2 **then**  (30) 将超过半数的标签作为新样本 synthSmpl的标签集合  (31)  **Return** synthSmpl  (32) **End function** |

## 多标签双重自适应采样算法

### 算法概述

已有的解决多标签不均衡性的方法要么是欠采样，要么是过采样，这种单方面采样来均衡数据集的方法虽然能够在一定程度上解决数据不均衡问题，但是采样过度可能导致数据集发生偏移，大类样本变成小类样本造成的大量信息丢失，小类样本由于产生的新样本过多而造成的过拟合问题。另一方面，目前的多标签采样算法LP和ML通过人工设定采样率来决定要删除的大样本数目，或者增加新的小类样本的数目，这种方法使用起来，首先得了解多标签数据集的分布情况，而多标签数据集本身涉及到一个样本拥有多个类标签的这种复杂特性增加了分析难度，粗略的设置采样率存在随机性大且未能考虑到整体数据集的分布情况，造成最终的数据集分布存在很大不确定性，使模型分类效果受到很大影响。

为了对以上问题进行处理，本节提出了多标签双重自适应重采样算法ML-DARS算法。该算法主要考虑到多标签数据集内部很大的差异性，将欠采样和过采样过程相结合，同时最小化人工干预对数据集造成的极大不确定性。该算法解决问题主要分为以下步骤：

* 1. 采样算法选择，本算法鉴于欠采样和过采样结合的思想，首先需要确定有效的欠采样算法、过采样算法为整个过程提供更好的基础。
  2. 小类标签、大类标签的划分，即选择哪些标签作为小类标签，哪类标签需要划分到大类标签集合。
  3. 如何确定采样过程结束，或者是如何确定数据集分布达到了较为均衡的情况。上述已讨论过，已有算法通过人为设定采样率规定需要减少或增加的样本数目存在极大的不合理性，而本算法充分利用数据集自身的内在分布信息来确定数据集的均衡性，进而作为采样结束的标准，获得更为均衡且合理的数据集。
  4. 如何集成欠采样和过采样两种算法来获得更为均衡的数据集，即具体的ML-DARS算法描述。

### 采样算法选取

正如上文介绍，目前最为流行的采样算法基于LP和基于ML的方法，LP方法带来标签集转化后产生的多分类标签规模大且数据量极不均衡等问题，基于ML方法的采样算法更为合理的均衡了数据集的分布。其中ML-RUS算法受LP-RUS算法思想的启发，对每个标签首先进行评估其不均衡度并合理的划分大小类标签，然后删除采样率所规定的样本数目，这些样本对应的标签不含有所有小类标签，既保证了小类标签对应的样本信息不会丢失，同时又删除了大类样本带来的冗余信息。该算法具体描述如4.2.2小节。

ML-SMOTE算法既考虑了ML-ROS算法的优势，即复制小类样本增加小类标签的样本数目，又解决了ML-ROS算法存在的重复样本导致的模型过拟合问题，对小类样本插值产生新样本，既有效解决了不均衡问题，又获得了合理且较好的模型分类效果，实际上， ML-SMOTE普遍代替ML-ROS算法用来解决多标签不均衡性问题。算法具体描述过程见4.2.2小节。

基于上述分析，ML-DARS算法将充分利用ML-RUS 算法思想和ML-SMOTE算法思想的优势，既解决大样本信息冗余问题，又解决了小样本缺乏信息表示问题， 极大均衡了多标签数据集的分布。

### 标签集合的划分

由于ML-DARS算法将欠采样过程和过采样相结合，算法开始之前需要确定哪些标签属于小类标签，哪些标签属于大类标签，然后根据划分好的大类标签集合和小类标签集合进行后续的欠采样和过采样过程。为了充分利用ML-RUS算法和ML-SMOTE算法的优势，ML-DARS算法同样根据标签间的不均衡程度对标签集合进行划分。根据第二章中的公式（2-12）、（2-13）当标签的不均衡度，将标签归为小类标签，否则将标签归为大类标签，同时一个样本可能既包含大类标签又有小类标签，对此，以尽可能保留所有小类标签的原始信息的思想，将该样本归为小类标签所在范畴，即保留该标签。随着采样过程的不断进行，需要不断更新数据集对应的不均衡度、，来评估数据集当前的状态，进而决定是否要继续采样。



### 数据集均衡性标准

上文提到已有的多标签重采样算法，人工设定采样率来决定采样是否结束，存在的问题是采样完成后的数据集可能并未达到均衡程度，虽然可以通过不断调整采样率来接近于数据的均衡程度，但随机性大且没有明确规定数据集是否达到均衡性标准。为了解决人工采样率存在的问题以及能够经过采样后数据集自适应性的达到均衡程度，ML-DARS算法进行了修正。

受标签间不均衡度的度量指标的启发，一般来说，标签不均衡度越是接近于MeanIR，说明标签集整体的分布越均衡，当IRLbl(ym)=MeanIR时，标签ym对应的样本数为整个数据集的平均样本数meanInstances，度量指标如公式(4-1)所示,max 为多标签数据集中最大标签对应的样本数，即最大样本数。

(4-1)



很容易得出，多标签数据集经过划分后，大类标签集中每个标签对应的样本数目大于平均样本数meanInstances，小类标签集中的每个标签对应的样本数则小于平均样本数meanInstances。 在采样过程中，为防止大类样本采样过度变成小类样本，小类样本由于样本数目大量增加而变成大类样本，造成数据集分布的偏离，ML-DARS利用meanInstances保持原始数据集的分布，即大类样本欠采样后始终大于或等于meanInstances，小类标签过采样后始终小于meanInstances，ML-DARS算法就是不断缩小大类样本、小类样本与meanInstances的差距的过程，从而接近于更为均衡的多标签数据集。

此外，MeanIR衡量多标签数据集整体的不均衡度，而最大类标签的不均衡度为1，随着采样过程的继续，若多标签数据集越来越均衡，那么数据集平均不均衡度MeanIR呈不断减小的趋势，也就是说MeanIR越小，数据集越均衡，其值不小于1。为了获得较为均衡的数据集，ML-DARS算法每过一定采样间隔，更新MeanIR，和各标签不均衡度IRLbl, 为了防止采样过度，需提前停止采样过程，需要对MeanIR设置容错范围threshold，即公式如（4-2），当大类标签的不均衡度大于MeanIR-threshold， 该标签对应的样本数目达到均衡标准，不在对其欠采样，同理当小类标签的不均衡度小于MeanIR+threshold，不在对其过采样。直到所有标签对应的样本均达到均衡，采样过程结束。容错能力threshold需要手动设置，一般而言threshold=5最为合适。

(4-2)



### ML-DARS算法描述

ML-DARS算法伪代码见算法3，首先需要将原始的多标签数据集根据标签间的不均衡性度量指标划分为小类标签集合、大类标签集合，同时获得原始标签集中度量均衡性的平均样本数，然后算法开始了欠采样过程和过采样过程相结合的采样过程，具体是取原始多标签数据集中的一个样本，对该样本的标签集合进行分析：如果该标签集中存在小类标签集合中一个或多个标签，则认为该样本属于小类样本，利用ML-SMOTE新样本合成思想，第一步获取该样本中存在的每个小类标签的近邻，第二步根据每个小类标签以及它的近邻进行插值和标签投票的方式形成新的样本，第三步将新样本加入原始多标签数据集中；受ML-RUS思想的启发，如果该标签集中的标签仅出现在大类标签集合中，则删除该样本；此外，随着采样过程的继续，有的标签既不存在于小类标签集合也不存在于大类标签集合，这类标签对应的样本数目一定程度上达到了使数据集均衡的目的。对该样本分析完后，由于多标签数据集的分布发生了变化，所以需要重新评估小类标签集合和大类标签集合。这里分为二步走策略，第一步，判断小类标签集合中的每个标签对应的样本数目是否超过平均样本数meanInstances,若超过，从标签集合中删掉对应标签，同理删除大类标签集合中小于等于平均样本数的标签。这一步主要是防止数据采样过度而偏离原始的数据分布，第二步，更新小类标签集合和大类标签集合中每个标签对应的不均衡度IRLbl以及新的整个数据集的平均不均衡度MeanIR，从小类标签集合和大类标签集合中删除不均衡度IRLbl在平均不均衡度的容错范围内的标签，继续循环，直到小类标签集合和大类标签集合为空为止，均衡了数据集的分布。

算法3 ML-DARS

|  |
| --- |
| Input: 训练数据集D, 近邻数量k  Output: 采样后的新数据集Y  (1) **for** each label **in** L **do**  (2) IRLbllabel =calculateIRLbl(D,label)  (3) **If** IRLbllabel < meanIR  **then**  (4) Label 加入maxBag 集合中  (5)  **else** label 加入 minBag中  (6) **End for**  (7) meanInstances =  (8) **while** (!minBag.isEmpty() or !maxBag.isEmpty() ) and |D|>0  (9) instance = getInstance(D)  (10) **If** instance 中出现了minBag中的小类标签 **then**  (11) New\_sample = ML-SMOTE(instance, minBag) //新样本合成  (12)  **End if**  (13) **If** instance 中除了大类标签外不存在其他标签 **then**  (14) Delete instance  (15)  **End if**  (16) 更新采样后的数据集中IRLbl， MeanIR  (17) 更新minBag, maxBag中的标签  (18) **For** each maxlabel **in** maxBag **do**  (19)  **If** maxlabel 的样本数目小于等于meanInstances  **then**  (20) Remove maxlabel from maxBag  (21)  **End if**  (22)  **End for**  (23) 类似(18)~(22) 步，将大于meanInstances 的标签从minBag中删除  (24)  **For** each minlabel **in** minBag **do**  (25)  **If**  IRLblminlabel <= meanIR+threshold  **then**  (26) Remove minlabel from minBag  (27)  **End if**  (28)  **End for**  (29) 类似（24）~(28)步， 将IRLbl大于meanIR-threshold的标签从maxBag中删除  (30) **End while** |

## ML-DARS算法实验

### 实验数据集

为了展现ML-DARS算法的优势，本次实验在心血管疾病数据集和2个公共多标签数据集进行了实验验证，其中包括了均衡与不均衡的数据集，并与ML-RUS、ML-SMOTE算法进行对比分析，具体信息见表4-1，其展示了原始数据集的样本数目、特征数目、标签数目、对应的标签基数、最大不均衡度、平均不均衡度、平均样本数度量指标等，一定程度上反映了数据的分布情况。可以看出Cardiovascular（心血管疾病数据集）标签基数很小，最大不均衡度达到了100以上，样本数目在10万以上，而平均样本数目才1779，说明数据分布很不均衡，而Chronic相对均衡，Emotions较为均衡。

表4-1 数据集分布指标

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Dataset* | *samples* | *features* | *labels* | | *card* | *MaxIRLbl* | *MeanIR* | | *meanInstances* |
| Cardiovascular | 119066 | 296 | 9 | 1.46 | | 121.49 | 54.37 | 1779 | |
| Chronic | 6591 | 310 | 10 | 2.38 | | 13.03 | 4.86 | 847 | |
| Emotions | 391 | 72 | 6 | 1.81 | | 1.77 | 1.49 | 112 | |

### 实验设置

本次实验基于开源的多标签学习库Mulan来进行开发， Mulan中包含大量的、成熟的多标签学习算法，同时也提供了多种公共多标签数据集，为多标签学习领域的研究提供了极大的便利。

本次实验利用Hamming-Loss 、SubAccuracy、F-measuresmacro、F-measuresmicro作为算法性能评价指标,其指标介绍见2.2.2节。F-measuresmacro先求各标签的F值再对所有标签取平均，F-measuresmicro先取所有标签各指标之和，然后在所有标签的整体上求F值，两者结合起来从宏观、微观两方面度量多标签分类的准确度。

为了评估多标签均衡性算法对分类效果产生的不同影响，本次实验选取了常用的多标签分类算法BR、ML-KNN、HOMER对采样后的数据集进行分类学习，并以J48、SVM、logistic算法作为基分类器，同时，本次实验采用ML-RUS、ML-SMOTE算法与ML-DARS算法作对比，其中涉及到求近邻算法，其k值均为5，通过不同的性能评价指标来评价算法的好坏。

### 实验结果分析

为了验证ML-DARS算法的有效性，首先对心血管疾病数据集利用算法ML-RUS、ML-SMOTE、ML-DARS算法进行了重采样，通过不同的分类器得到最终的结果如表4-2、表4-3所示， ML-DARS算法得到的结果普遍优于ML-RUS和ML-SMOTE，特别是在SubAccuracy指标上十分明显，在BR-SVM上的F-measuremacro表现稍差，而ML-RUS算法的表现效果整体要略优于ML-SMOTE算法，也就是说直接采用ML-SMOTE算法不能有效解决心血管疾病数据集中的类别不均衡问题。

针对公共数据集，如图4-1、4-2、4-3所示，ML-DARS算法在Hamming-Loss上总体小于ML-RUS、ML-SMOTE算法， 由于chronic 、emotions数据集本身就较为均衡，所以ML-DARS算法的表现并不明显，但依然可以看出该算法与其他两种算法不分高低，除了对于F-measuremicro指标，ML-DARS算法在emotions数据集上表现较ML-SMOTE算法差一点，其他方面表现都略高。

表4-2 ML-DARS算法在心血管疾病数据集性能比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier Algorithm | Hamming-Loss | | | SubAccuracy | | |
| ML-RUS | ML-SMOTE | ML-DARS | ML-RUS | ML-SMOTE | ML-DARS |
| BR-Logistic | 0.0859 | 0.0900 | 0.0675 | 0.4639 | 0.4424 | 0.5747 |
| BR-J48 | 0.1518 | 0.1532 | 0.0743 | 0.2425 | 0.2184 | 0.5440 |
| BR-SVM | 0.0787 | 0.0848 | 0.0674 | 0.4837 | 0.4541 | 0.5695 |
| HOMER-j48 | 0.1909 | 0.1878 | 0.0854 | 0.1638 | 0.1600 | 0.5076 |
| MLKNN | 0.1082 | 0.1471 | 0.0833 | 0.3664 | 0.2401 | 0.5147 |

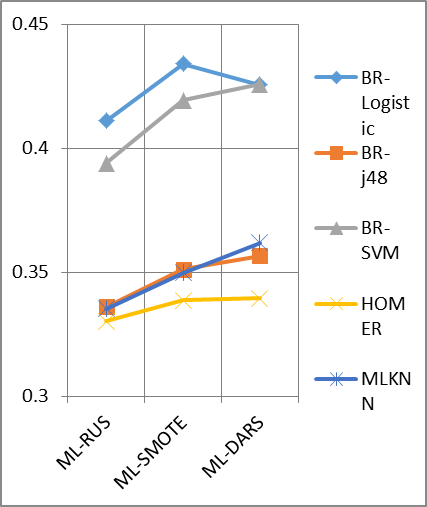
表4-3 ML-DARS算法在心血管疾病数据集性能比较

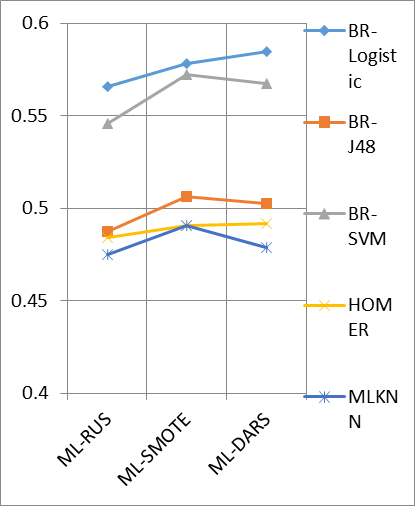
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier Algorithm | F-measuremicro | | | F-measuremacro | | |
| ML-RUS | ML-SMOTE | ML-DARS | ML-RUS | ML-SMOTE | ML-DARS |
| BR-Logistic | 0.7506 | 0.7427 | 0.7817 | 0.3377 | 0.3174 | 0.3250 |
| BR-j48 | 0.6282 | 0.6256 | 0.7742 | 0.3013 | 0.3016 | 0.3983 |
| BR-SVM | 0.7664 | 0.7555 | 0.7802 | 0.3158 | 0.3145 | 0.3098 |
| HOMER-j48 | 0.5890 | 0.5871 | 0.7550 | 0.2888 | 0.2864 | 0.4062 |
| MLKNN | 0.6910 | 0.6178 | 0.7345 | 0.2428 | 0.2373 | 0.2837 |



1. Chronic数据集 b) emotions数据集

图4-1 Hamming-Loss 性能比较





a) Chronic数据集 b) emotions数据集

图4-2 Accuracy 性能比较



a)Chronic 数据集 b)emotions 数据集

图4-3 F-measuremicro性能比较

## 本章小结

本章对多标签数据集不均衡性问题展开讨论，详细介绍了现有的且常用的多标签不均衡性方法ML-RUS、ML-SMOTE，并给出了算法伪代码。在此基础上，为了解决现有的多标签不均衡算法带来的大类样本采样过度造成的信息丢失而小类样本过采样造成的信息冗余，本章提出了一种多标签双重自适应采样算法ML-DARS，详细介绍了该算法的原理，并利用实际数据集进行了实验验证，与ML-RUS、ML-SMOTE算法进行对比分析，证明了ML-DARS算法能够获得更为均衡的数据集。

# 基于混合策略的心血管疾病预测模型

## 引言

目前已有的多标签分类算法在一定程度上很好的解决了不同应用场景下的分类预测问题，而心血管疾病数据集存在两大特性：一是症状相似性，即患有相同心血管疾病的患者一般有相同的表现症状，例如所有的高血压患者一般会出现头晕、胸闷、心悸等症状，二是心血管疾病之间存在很强的互相影响关系，例如高血压患者更容易患冠心病、心衰等大多数易危及生命的心血管疾病。

基于近邻的标签信息来评估样本的隶属度，这种方法利用局部信息构造决策面，通过度量相似症状的患者患有的疾病来确定待预测患者可能患有的心血管疾病，但是这种方法缺乏表示标签间的相关信息，而且有限的局部信息也会影响分类器的分类性能。而目前充分利用标签间的相关信息的分类算法纵观数据集的全局信息，势必会忽略局部细节。

为了充分挖掘心血管疾病数据集信息获得更好的预测效果，考虑到ML-KNN算法能够根据近邻样本信息来预测待测样本，有效的利用了心血管疾病数据中症状相似的特性。而RAKEL算法既解决了LP计算复杂度问题，又充分了标签间的相关性特点，是心血管疾病间相互关系挖掘的一大助力。基于此，本文结合ML-KNN和RAKEL算法的优势，提出了基于混合策略的多标签分类模型。

## 基于混合策略的心血管疾病预测模型

### 大规模数据分批训练策略

众所周知，RAKEL算法是对LP所带来的大规模标签组合、数据极不均衡性问题的改进，在一定程度上降低了LP问题造成的极大的时空复杂度。然而，RAKEL算法的本质依然是基于LP思想，即便通过标签子集的方式减少了标签组合的粒度进而降低了一次训练的规模，但训练样本总体的数目并未减少，而且该算法需要为每个标签子集训练一个分类模型，也会占用大量的空间资源，经常出现内存空间不足而溢出的现象，在大规模训练集面前，单凭RAKEL算法是无法有效训练数据的。

基于此，本文对心血管疾病数据集进行划分，在RAKEL算法能够处理的数据规模范围内，将心血管疾病数据集分成多个子集，对每个子集训练一个RAKEL模型，在测试集上进行预测，获得测试样本的一次预测结果，循环多个子集，对测试集上样本的多次预测结果取均值，作为RAKEL训练过程中的测试样本预测结果。在这里划分数据子集并非随机，数据集的规模越大，目标标签对应的样本数目差异越大，数据越是不均衡，同样为了确保划分后的子集更为均衡，数据子集划分策略如下：

首先，心血管疾病数据集按照小节2.3.2的划分标准，获得小类标签集合和大类标签集合，出现在小类标签集合的所有样本被看成是小类样本集合Dmin，其他的样本组成大类样本集合Dmax。

然后，通过对心血管疾病数据集的试探，获得RAKEL算法大约一次能够处理的样本数目，记为 ，即获得数据子集的大约估计。由于小类样本集中存在大类标签，而大类样本集中不存在小类标签，为得到更为均衡的数据集，将数据子集中的小类样本和大类样本比设为1.5:1到2:1之间，即均衡了数据的分布，保持了原有数据分布的总体情况。 据此计算，可得到数据子集中需要的小类样本数目 和大类样本数目 。

最后，根据小类样本集合Dmin的大小和数据子集中需要的小类样本数目 ，可将小类样本集合约均分为k1份，同理大类样本集合均分为k2份，对k1份小类样本子集分别与k2份大类样本子集进行混洗，每次使用一次混洗结果进行RAKEL算法训练，共混洗k1\*k2次，即为训练子集的数目。

这样，既解决了RAKEL算法在大规模训练样本集上的障碍，又保证了更为均衡的训练样本集，在多个不同的样本子集上训练而保证了和大规模样本情况下一样的预测效果。

### 模型构建思想

根据心血管疾病数据集的特性，本文提出了基于混合策略的多标签学习方法（MR）来构建心血管疾病预测模型。该模型分为三个步骤如下：

首先利用ML-KNN算法获得训练集上每个标签被误分类的概率，作为衡量该标签能够被局部信息所表示的能力，由于多标签数据集存在极不均衡性，即多个标签组成的数据集很大，而一种标签对应的样本数目占总样本数的比例则非常小，多标签分类算法在训练过程中，待预测标签对应的样本集为正类，其他样本对应于负类，预测结果通常偏向于大类，为了准确反映待预测标签的被误分类的程度，利用正类样本中被预测为负类的样本所占比例作为每个标签被误分类的概率，记为α，公式如（4-3）所示，表示标签被ML-KNN误分类的概率。

 (4-3)

然后在此基础上，对每一个测试样本，经过ML-KNN预测得到对应标签的预测结果记为。

利用5.2.1大规模数据分批训练的思想，以RAKEL算法为基准训练样本，获得一次的预测结果记为。

最后加权两种算法的预测结果作为最终的预测值，加权结果如公式（4-4）表示。

 (4-4)

其中。 待预测样本通过阈值求出，见公式（4-5），一般。

 (4-5)

该过程获取MLKNN算法对每个标签的误分类率、对测试数据的预测结果和RAKEL算法的训练、预测过程可并行进行，弥补了混合策略所带来极大的时间开销。

基于混合策略的多标签学习模型MR伪代码描述如下

算法4 MR算法伪代码

|  |
| --- |
| Input: 训练数据集D, 测试数据集T  Output: 预测结果，模型评估结果  (1) L ← 为训练集D中对应的原始标签集合  (2) 𝛂 ← Hybrid parameter  (3) N← number of misclassified samples about every label  (4) M ← number of instances about every label  (5) ML-KNN train dataset D  (6) **For** each instance **in** D **do**  (7) Truth ← true label of instance  (8) Predict ← ML-KNN predict instance  (9)  **For** each label **in** L  (10) **If** truthlabel ==1 and predictlabel ==0  **then**  (11) Nlabel .add(1)  (12) **End if**  (13)  **End for**  (14) **End for**  (15) **For** each label **in** L  (16) 𝛂label ← Nlabel /Mlabel  (17) **End for**  (18) Double[][] rakelPrediction ← integrateRAKEL(D, T) //将数据集划分并分别训练，返回所有分类器对测试样本预测均值  (19) predictFinal ← finally Hybrid result for every sample  (20) **for** each instance **in** T **do**  (21) predictmlknn ←MLKNN .predict(instance)  (22) predictrakel ← rakelPrediction[instance.location]  (23) **For** each label **in** L  (24) predictFinallabel ← (1-𝛂label) predictmlknn + αlabel predictrakel  (25)  **End for**  (26) **End for**  (27) Evaluation of prediction results |

数据集划分为大类子数据集和小类子数据集，大小类标签数据集混洗，RAKEL算法分批训练，测试样本中每个标签的值取所有RAKEL模型的预测均值，数据划分混洗算法伪代码描述如下。

算法5 RAKEL分批训练数据集

|  |
| --- |
| Input: 训练数据集D, 测试数据集T  Output: 所有测试样本最终的预测结果  (1) Function integrateRAKEL( D , T)  (2)  **for** each label **in** L **do**  (3) IRLbllabel =calculateIRLbl(D,label)  (4)  **If** IRLbllabel < meanIR **then**  (5) Label 加入maxBag 集合中  (6)  **else**  label 加入 minBag中  (7) **End for**  (8) smallSamples ← 训练集中出现在minBag集合中的所有样本  (9) maxSamples ← 训练集中出现在maxBag集合中的所有样本  (10) smallSampleSplit[] ← splitInstances(smallSamples , k1);//小样本集划分为k1份  (11) maxSampleSplit[] ← splitInstances(maxSample, k2); //大样本集划分为k2份  (12) double[][] val ← 所有测试样本最终的预测结果  (13) **For** each smallSamples **in** smallSampleSplit **do**  (14) **For** each maxSamples **in** maxSampleSplit **do**  (15) trainingSet ← mergeSamples(smallSamples, maxSamples)  (16) rakelmodel← RAKEL. train (trainingSet);  (17) **For** each sample **in** T **do**  (18) output← rakelmodel.prediction(sample)  (19) Val +=output的结果按在val的位置进行累加  (20)  **End for**  (21)  **End for**  (22)  **End for**  (23) val[i][j]=val[i][j]/(smallSampleSplit.length\*maxSampleSplit.length)  (24) **return val**  (25) **End function** |

## 实验

### 数据集分析与处理

本次实验在4.2节基础上进行的，对原始心血管疾病数据集通过ML-DARS算法进行处理，降低冗余数据，调整数据分布，获得合理的数据集。

原始数据集基数card为1.46，密度density为0.16，共存在197种不同的标签组合，数据分布如图4-4所示，由于心血管疾病数据集有9个目标标签，所以理论上组合情况一共有512种，这里我们通过对原始数据集标签规模排序，取出了40种标签组合的分布情况。经过ML-DARS算法处理，新数据集基数card为2.24，密度density为0.25，共存在199种不同的标签组合，前40中标签组合分布如图4-5，可以看出，前5种标签组合占数据集的大部分，特别是标签组合Label1，在原始数据集中几乎占据了全部，经处理后，降低了该组合标签的比重，ML-DARS算法有效均衡了数据集的分布，且在总体上保持了原有数据分布。

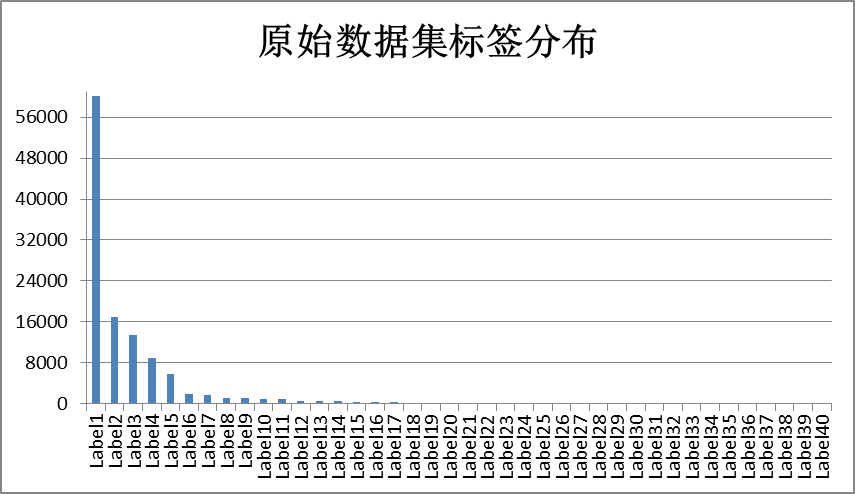


图4-4 原始数据集部分标签分布

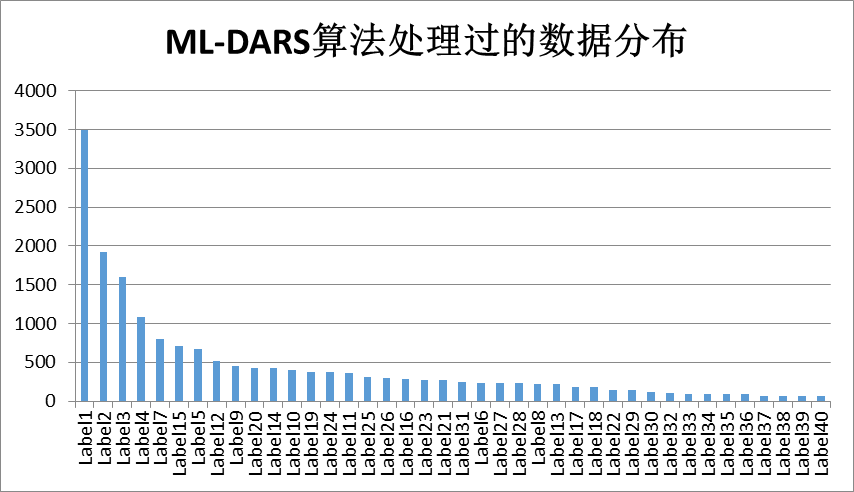


图4-5 ML-DARS算法均衡后的部分数据分布

经处理，将新的多标签数据集应用于模型预测过程。新数据集包括训练集和测试集数据统计如表（4-4）所示：

表4-4 最终的数据集统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **数据集** | **样本数** | **特征数** | **标签组合数** | **标签基数** |
| 训练集 | 19896 | 296 | 199 | 2.24 |
| 测试集 | 6682 | 296 | 112 | 1.48 |

### RAKEL算法分批训练策略

首先对心血管疾病数据集统计大小类样本占的比例，小类样本集大小为8294 ，大类样本集大小为11602。经过对RAKEL算法在心血管疾病数据集上的多次试探，在RAKEL算法尽可能充分学习到规律的情况下，需要将训练子集数目控制在7000到7200之间，超过该范围使用RAKEL算法在现有软硬件环境下发生内存溢出。

然后为了缓解数据不均衡现象，初步思路将划分后的训练子集大小样本比例尽可能接近1.5:1，即训练子集中小类样本数目约需要4200到4320之间，大类样本约需要2800到2880左右。之后考虑到现有样本的实际情况，由此，对现有的8294小类样本集将其划分为2部分，每部分为4147个小类样本，而大类样本11602则划分为4部分，每部分约为2900个大类样本，最后每个小类样本子集分别与每个大类样本子集进行混洗，得到8份数据集分别用于RAKEL算法训练预测，获得8个RAKEL模型，每个模型对测试样本进行预测，最终的预测结果取8个模型预测结果的均值作为此部分的预测结果。

### 分类结果

本次基于混合策略的多标签分类框架MR首先涉及到了ML-KNN算法， 利用该算法获得每个标签对应的误分类率，作为与RAKEL算法预测结果的加权系数。这里所有算法涉及到基分类器都默认为J48，ML-KNN中近邻系数k采用算法包中默认值10，获得的权重系数误分率如表4-5 所示，可以看出大多数标签由于对应的样本比例非常小，利用ML-KNN算法小类标签被误分的概率几乎都在50%以上，极大的降低了模型整体的性能。利用RAKEL算法时，将子集大小设置为3，共训练了18个RAKEL分类器。这里为了更明显的表明混合策略的有效性，分别取基于混合策略分类模型MR-J48、ML-KNN和RAKEL-J48单独分类所得的Macro-averaged F-measure结果，如图4-6所示，该结果更为细节的展示了每个标签分类精度上的变化，从整体上看，可以看出MR-J48中权重系数误分率的影响下， MR方法结合了ML-KNN和RAKEL算法的优势，获得了更好的分类精度，特别是小类标签心衰疾病和肾衰竭疾病的精确度明显优于ML-KNN和RAKEL单独分类的结果，误分类权重很大的脑卒中、心肌缺血、心脏功能病变等疾病的精确度在RAKEL算法的影响下略有提升。

表4-5 权重系数误分率

|  |  |  |
| --- | --- | --- |
| *目标标签* | *权重（误分率）* | *标签大约所占样本比例* |
| cerebral\_apoplex（脑卒中） | 0.9157 | 17% |
| hf\_ICD（心衰） | 0.6037 | 8% |
| cardiac\_infarction（心肌梗死） | 0.6308 | 8% |
| Nephropathy（肾衰竭） | 0.6103 | 11% |
| ischemia\_myocardial（心肌缺血） | 0.8307 | 8% |
| cardiac\_function\_lesion（心脏功能病变） | 0.9898 | 11% |
| coronary\_heart\_disease（冠心病） | 0.4345 | 39% |
| Hypertension（高血压） | 0.0526 | 80% |
| Diabetes（糖尿病） | 0.3881 | 42% |

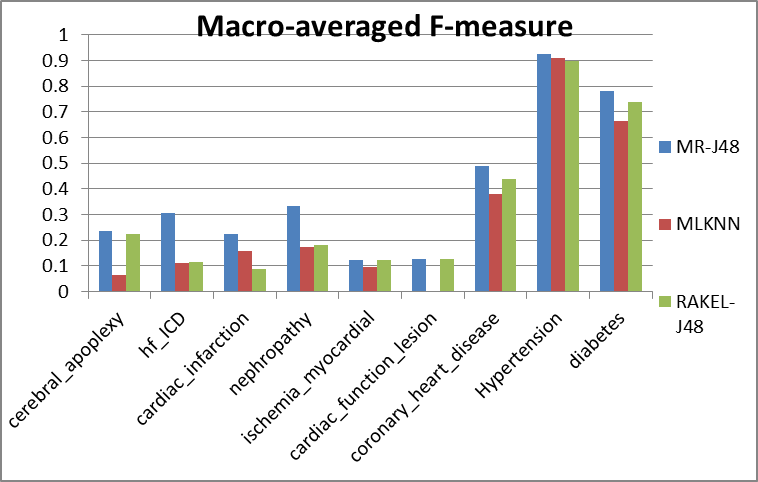


图4-6 Macro-averaged Precision结果

为了证明混合策略框架MR的有效性，本文其与其它现有的多标签分类算法进行了详细的比较，首先对比MR框架中用到的ML-KNN、RAKEL算法，下面从Hamming-Loss、排序损失RankingLoss、SubsetAccuracy、Micro-averaged F-measure等常用的评价指标展示了MR的分类效果。如图4-7、4-8、4-9、4-10所示，可以看出，基分类器为J48的情况下，MR在Hamming-Loss、RankingLoss均要低于其他算法，在准确度方面，MR方法的SubsetAccuracy、Micro-averaged F-measure均要高于其他算法，因此，基于MR框架的心血管疾病预测模型获得了更有效更好的预测效果。

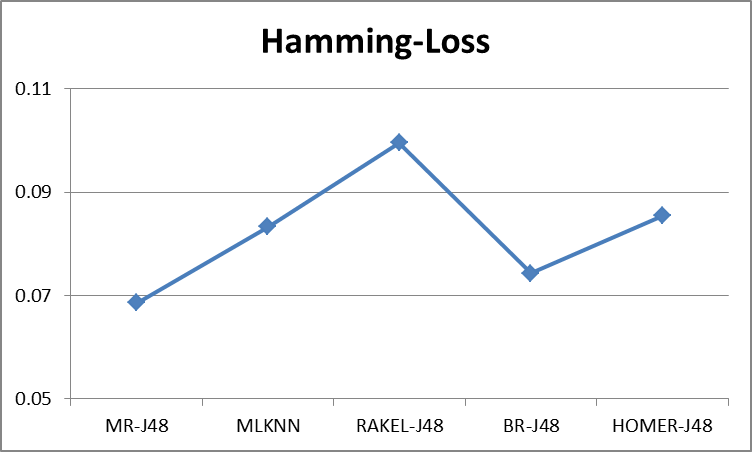


图4-7 Hamming-Loss

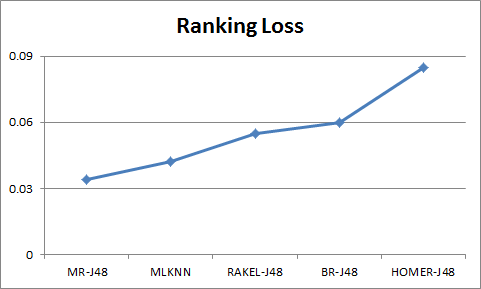


图4-8 RankingLoss

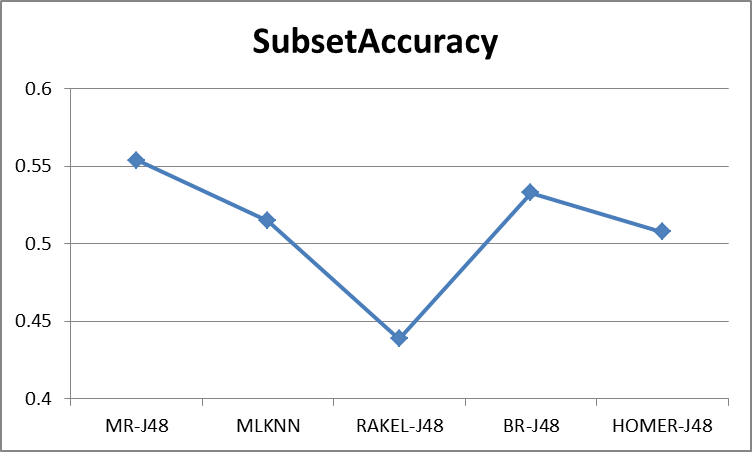


图4-9 SubsetAccuracy

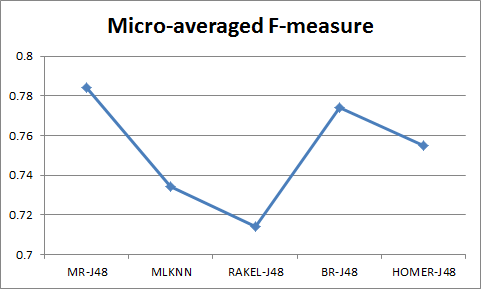


图4-10 Micro-averaged F-measure

表4-6、4-7详细记录了各分类器对心血管疾病数据集的分类效果， 从表中可以看出，除了Macro-averaged F-measure指标中MR比BR-J48、HOMER算法较低外，其他算法的分类效果均在MR之下。

表4-6 分类结果对比1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier Algorithm** | Hamming-Loss↓ | SubAccuracy↑ | Micro-averaged F-measure↑ | Macro-averaged F-measure↑ |
| **MR-J48** | **0.0686** | **0.5538** | **0.7844** | 0.3636 |
| **BR-J48** | 0.0743 | 0.544 | 0.7742 | 0.3983 |
| **MLKNN** | 0.0833 | 0.5147 | 0.7345 | 0.2837 |
| **HOMER-J48** | 0.0854 | 0.5076 | 0.755 | **0.4062** |
| **RAKEL-J48** | 0.0995 | 0.4386 | 0.7143 | 0.3264 |

表4-7分类结果对比2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier Algorithm** | Ranking Loss↓ | Coverage↓ | Average Precision↑ | OneError↓ |
| **MR-J48** | **0.0342** | **0.8661** | **0.9293** | **0.0815** |
| **MLKNN** | 0.0422 | 0.9446 | 0.9132 | 0.0942 |
| **RAKEL-J48** | 0.0548 | 1.0978 | 0.8924 | 0.1423 |
| **BR-J48** | 0.0599 | 1.1369 | 0.8966 | 0.1236 |
| **HOMER-J48** | 0.0847 | 1.4356 | 0.8661 | 0.1534 |

## 本章小结

本章作为心血管疾病预测模型中的最后一步，也是最重要的一步，首先在前面章节的基础上，为获得更好的分类效果，根据心血管疾病数据集的特性，本文采用基于混合策略的多标签学习框架，提出了ML-KNN算法和RAKEL算法相结合的多标签分类框架，建立心血管疾病预测模型，该模型对基分类器取长补短，并且在RAKEL算法无法处理大数据量的情况下，防止欠采样导致的信息丢失降低分类性能，同时获取更为均衡的数据集，对心血管疾病数据集按照均衡性指标进行划分，采用大类样本小类样本混洗的方式，反复多次，保证每个数据集中的每个样本参与训练，在此基础上进行了实验分析验证，结果表明该模型相比于现有的多标签分类算法取得了很好的预测效果。

# 结论与展望

## 总结

本文从医疗大数据项目出发，以心血管疾病预测为目标，从项目所提供真实的、可靠的医疗数据出发，通过统计分析确定研究对象，提取研究对象对应的数据，并进行预处理和有效的特征因素选择，经过这一系列的处理，获得了干净的、完善的具有可分类特性的心血管疾病数据集，此外，为了提高多标签分类的可靠性，对处理后的数据集进行了数据均衡性采样，根据心血管疾病数据集的多标签特性以及确保采样后的数据仍能保持原始数据的分布，提出了多标签双重自适应随机采样算法，并应用心血管疾病数据集和两种不同的公开数据集进行了实验验证，结果表明新的采样算法相比于现有的采样算法能够很好的解决多标签数据均衡性问题，并且在一定程度上提升分类效果，为之后的多标签分类预测打下了坚实的基础。在之前的基础上，本文利用现有多标签分类算法的优势，结合心血管疾病数据集的特性，提出了以ML-KNN和RAKEL算法为基础且基于混合策略的多标签分类框架，预测结果表明，利用多标签分类算法能够有效解决医学领域疾病预测这样一个复杂难题，为数据挖掘技术应用于医学数据研究提供了新方法、新思路。

## 展望

本文为解决心血管疾病预测问题，同时考虑到心血管疾病间存在的复杂关系，多标签分类技术被作为此次预测的手段。本文为了将多标签分类算法应用于心血管疾病预测问题，提出了一套完整的医疗数据处理流程， 获得了干净且真实的心血管疾病数据集，提出的均衡算法解决多标签数据集不均衡问题。 但本文的心血管疾病预测模型依然有待进一步提升，处理不均衡问题上，是否能够改进新的采样算法使得减少采样过程中标签间信息丢失问题，获取更为均衡的数据集，基于混合策略的模型虽然获得了更高的分类效果，但是训练多个分类器要花费更长的时间，能否降低该模型的预测时间，是下一步要做的工作。

# 致 谢

时间如白驹过隙，转眼间，三年的研究生生涯即将结束，这段日子中，我经历了许多，也成长了许多。在毕业论文撰写过程中，虽然遇到了各种大大小小的困难，但是老师、家人、同学还有实习中遇到的同事他们都给予我莫大的关怀和帮助，在此我衷心感谢他们为我所做的一切。

由衷感谢我的研究生导师魏恒义老师对我这三年来在学业上的耐心教导和帮助。工作中，魏老师认真严谨、兢兢业业，谆谆教导每一位学生，生活中魏老师更是为人和善、对待学生和蔼可亲，不论何时，我遇到任何问题，魏老师都能耐心解答，设身处地的为我们着想。在这里再次感谢魏老师对我毕业论文写作过程中无私的细心指导和建议，有魏老师的帮助，我的论文才能不断改进直到完善。此外，感谢实验室的师兄师姐在学习、生活中对我无微不至的热心帮助让我克服了许多困难。

感谢深圳中科院先进技术研究院的指导老师蔡云鹏还有一起参与这个项目的同学们，我很庆幸参与到医疗大数据这个项目中，第一次接触多标签分类问题，也切身体会了真正的大数据场景，通过对大量医疗数据处理，提高并拓展了我的专业技能，其中也遇到了很多障碍，一方面由于对医学知识缺乏，致使数据处理方面费时费力，幸好有先进院的老师同学的帮助，使得我的工作可以继续，另一方面在用大数据平台spark时，异常数据定位，虽然利用spark监控页面能够了解任务运行状况，调试程序，但是有些情况下数据处理的前面步骤未出现问题，到后面运行出错，致使不断向前追溯处理过程中的隐匿错误，通过老师同学们的耐心指导我充分利用各种开发工具、促使程序完善，最终形成了一个医疗数据处理的完整流程，为毕业论文顺利完成提供了重要保障。

最后，感谢我的父母，感谢他们一直以来对我的支持与鼓励，是我不断坚持与前进的力量。

# 参考文献

1. 王普．多标记学习算法研究及生物医学数据挖掘中的应用[D]．中科院深圳先进技术研究院，2017．
2. 李思豪，陈福才，黄瑞阳．一种多标签随机均衡采样算法[J]．计算机应用研究，2017， 34(10):2929-2932．
3. 肖雨奇．多标签学习应用于中医诊断帕金森中类别不均衡问题研究[D]．南京大学，2016．
4. 陈旭，刘鹏鹤,，孙毓忠等．面向不均衡医学数据集的疾病预测模型研究．Vol.40，在线出版号 No.155．
5. 何伟骏．基于层次—互斥模型的多标签分类算法的研究与应用[D]．中山大学，2015．
6. 王臻．基于学习标签相关性的多标签分类算法[D]．中国科学技术大学，2015．
7. 张居杰．多标签学习中关键问题研究[D]．西安电子科技大学，2016．
8. 谷炫志．基于情感的多标签个性化音乐分类技术的研究与实现[D]．浙江大学，2016．
9. 陈自洁．多标签分类问题的图结构描述及若干学习算法的研究[D]．华南理工大学，2015．
10. 李熙铭．基于主题模型的多标签文本分类和流文本数据建模若干问题研究[D]．吉林大学，2015．
11. 方铭．多标签分类算法的研究及其在中医诊断帕金森领域的应用[D]．南京大学，2015．
12. 雷少正．基于量表测评的帕金森用药推荐模型研究[D]．南京大学，2014．
13. 王刚．多标签学习及其在帕金森中医诊断中的应用[D]．南京大学，2014．
14. 周浩．中文多标签文本分类算法研究[D]．上海交通大学，2014．
15. 董纯洁．基于实例与逻辑回归的多标签分类模型[D]．南京大学，2013．
16. Tsoumakas G, Katakis I, Taniar D. Multi-Label Classification: An Overview[J]. International Journal of Data Warehousing & Mining, 2008, 3(3):1-13.
17. Wan S, Mak M W, Kung S Y. Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2016, 14(1):212-224.
18. Hofer T, Hofer T, Schumacher M, et al. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases[J]. Computers in Biology & Medicine, 2015, 65(C):34-43.
19. Zhang W, Liu F, Luo L, et al. Predicting drug side effects by multi-label learning and ensemble learning.[J]. Bmc Bioinformatics, 2015, 16(1):365.
20. Zhang M L, Zhou Z H. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8):1819-1837.
21. Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification ☆[J]. Pattern Recognition, 2004, 37(9):1757-1771.
22. Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3):333-359.
23. Read J. A pruned problem transformation method for multi-label classification[C]// Proc. 2008 New Zealand Computer Science Research Student Conference. 2008:143--150.
24. Tsoumakas G, Katakis I, Vlahavas I. Effective and efficient multilabel classification in domains with large number of labels[J]. Ecml/pkdd Workshop on Mining Multidimensional Data, 2008.
25. Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[M]. Elsevier Science Inc. 2007.
26. Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data[J]. 2009:667-685.
27. Schapire R E, Singer Y. BoosTexter: A Boosting-based System for Text Categorization[J]. Machine Learning, 2000, 39(2-3):135-168.
28. He J, Gu H, Liu W. Imbalanced Multi-Modal Multi-Label Learning for Subcellular Localization Prediction of Human Proteins with Both Single and Multiple Sites[J]. Plos One, 2012, 7(6):e37155.
29. Zhang M L. Ml-rbf: RBF Neural Networks for Multi-Label Learning[M]. Kluwer Academic Publishers, 2009.
30. Tahir M A, Kittler J, Bouridane A. Multilabel classification using heterogeneous ensemble of multi-label classifiers[J]. Pattern Recognition Letters, 2012, 33(5):513-523.
31. Seiffert C, Khoshgoftaar T M, Hulse J V, et al. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance[J]. IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans, 2010, 40(1):185-197. 5(C):34-43.
32. Anand Borad, Healthcare and Machine Learning: The Future with Possibilities, https://www.einfochips.com/blog/healthcare-and-machine-learning-the-future-with-possibilities/ .

# 攻读学位期间取得的研究成果

[1] 蔡云鹏，李奇，李晓燕，杨玉洁，程敬，吴红艳，李烨．一种数据处理平台和系统．申请号: 201710245078.3 .

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)