

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）：专业硕士（计算机）

2018年1月

**Research and application of prediction model of cardiovascular disease based on multi-label classification**

**By**

**Jing cheng**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

Supervisor: S.E.Hengyi Wei

Software Engineering

January 2018

**声 明**

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知,除了文中特别加以标注和致谢的地方外,本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名: 日期:

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问题，一种基因多种功能，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，在对心血管疾病问题、多标签分类算法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，选择效果最好的特征集。第四步为解决多标签数据集分布的稀疏性、不均衡性，针对多标签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。第五步将多种多标签分类算法应用于心血管疾病预测并作比较。

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、多标签数据集严重不均衡、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。

**关键词**： 心血管疾病， 多标签， 多标签不均衡性，重采样

[[1]](#footnote-1)

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 make an extensive and profound research on the issue of cardiovascular disease and multi-label classification algorithm. The first step understand the meaning of data such as data attributes, some of the most common cardiovascular disease. analysing data distribution, determining the object of study. The second steps extract 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. The fifth step is to apply a variety of multi-label classification algorithms to the prediction and comparison of cardiovascular diseases.

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 balance the distribution of multi-label data sets and improve reliability and accuracy model.

Key words: cardiovascular disease, multi-label, imbalanced dataset, resampling

**目 录**

1. 绪论

1.1 课题研究背景和意义

1.2 国内外研究现状

1.2.1 多标签学习的发展现状

1.2.2 机器学习在生物医学领域的发展现状

1.3 主要研究内容及创新点

1.3.1 本文主要研究内容

1.3.2 创新点

1.4 课题来源及论文结构

1. 多标签学习理论及典型方法介绍

2.1 引言

2.2 多标签学习的形式化描述

2.3 多标签学习的评价指标

2.4 多标签学习的工作原理及典型方法

2.4.1 问题转化方法

2.4.2 算法适应方法

2.5 本章小结

第三章 多标签学习算法在心血管病预测中的应用

3.1 模型建立的目标和步骤

3.2 研究对象的判定与抽取

3.3 样本数据预处理方法

3.4 数据特征选择

3.5心血管疾病预测模型

3.5.1 多标签数据集分布评估

3.5.2 建立多标签预测模型

3.5.3 模型预测能力测试

3.6 本章小结

1. 实验及结果分析

4.1 多标签数据集

4.2 预测模型的特征选择实验

4.3 多标签心血管疾病预测实验

4.4 实验环境

4.5 实验数据

4.6 结果分析

4.7 本章小结

1. 结论与展望

5.1 结论

5.2 展望

致 谢

参考文献

第一章 绪论

# 1.1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、体检信息等医疗数据，还涉及到公共卫生管理信息特别是疾控部门，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。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 心血管疾病所占比例

监督学习作为机器学习领域中最多被研究、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重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，在现代医疗的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的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，多标签分类技术应用医学领域依然存在很大挑战。

# 1.2国内外研究现状

## 1.2.1多标签学习的发展现状

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。经过多年来学者们不断地深入研究，多标签分类问题有了许多显著地解决方案，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）。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。AA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等人提出了用于多标签数据[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]，成为机器学习中重要分支。

## 1.2.2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。

# 1.3主要研究内容

本文用于研究的数据来源为各医疗机构的真实且保密的门诊数据，数据量达到了千万级，因此本文用于研究的数据十分宝贵，也为心血管疾病数据在多标签学习领域的实践打下了十分重要的基础。本文的主要目标是对实际医学数据经过处理，通过修正现有的多标签重采样算法，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，应用多标签分类算法建立心血管疾病预测模型中，提高了分类效果，主要研究工作包含以下几个方面：

（1）认识了解心血管疾病，分析研究大数据技术、机器学习、多标签分类算法和统计评价数据集的方法，数据集不均衡性问题。调研国内外基于机器学习、多标签分类算法应用于疾病预测的案例。

（2）通过研究统计分析实验数据，本文对心血管疾病预测模型进行详细的需求分析，建立了需求模型，其过程包括数据加载、确定并抽取研究对象、数据预处理、特征选择、模型建立、模型评价等功能的活动流程。

（3）模型设计了系统的体系结构，共包括4层，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：

数据采集：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于Oracle数据库中，将其心血管疾病数据从Oracle 导出并加载到hdfs中存储，利用医学知识、spark平台和ETL技术，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。

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。

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。

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，为此将多种多标签算法应用于心血管疾病预测模型，评价在当前数据集上的表现。

模型结果测试和改进：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，深入分析不同的重采样算法导致的多标签分类效果的差异性，比较模型预测结果显示，新的重采样算法ML-DARS得到了更加均衡的数据集，获得更好的分类效果。

# 1.4 课题来源及论文结构

本文主要工作分为三大部分：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；三是本文对用于研究的心血管疾病数据从提取，利用医学知识和数据预处理技术清洗数据，到特征选择，最终的心血管疾病数据应用到多标签分类算法中。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，多种的模型评价指标，现有的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，同时介绍了评价多标签数据集分布的指标以及不均衡度。

第三章 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多标签数据集，并在此基础上与现有的采样算法进行实验对比分析。

第四章 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，最后将其应用于多标签分类算法等全程的预测模型建立过程。

第五章 对本文工作进行总结与评价，指出了研究中存在的不足，以及下一步研究进展。

第二章 多标签类别不均衡性及分类方法介绍

# 2.1 引言

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。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

分类模型

特征表示

特征选择

训练样本

预测结果

图 2 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。

[14]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。

# 2.2 多标签学习介绍

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。

## 2.2.1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：，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：。另一种常用的多标签测量方法为标签差异性：，样本空间中出现的不同标签集的数量。

## 2.2.2 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)。

**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microaveraged 、macroaveraged。在这里可又分为基于样本的评估和基于标签的评估方法。

* 基于样本的评估方法

Hamming loss的定义如下：

 （1）

∆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：

 （2）

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：

  （3）

* 基于标签的评估方法

Microaveraging和macroaveraging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。Microaveraging和macroaveraging的计算方法如下：



 （4）

这些指标越大越好，最优值是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的样本所占比例：

 （5）

其中



Coverage指标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。

 （6）

Ranking loss：

 （7）

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。

## 2.2.3 多标签学习的典型处理方法

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。

**基于问题转换的方法**

* 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。如下表，表1代表原始数据集，转换成BR算法要求的数据集如表2，将表1数据集划分为四个不同的二分类数据集。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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1 原始多标签数据集

|  |  |  |
| --- | --- | --- |
| 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 |

表2 BR转化后的数据集

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，出现数据不均衡性。

* 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3， LP中新类别的生成。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

表3 LP转换后的新类别

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。

* 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

（8）

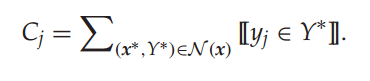
表示标签在k个标签子集中实际出现的频率，表示标签

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。

**算法适应性方法**

* ML-KNN

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，公式如下：

（9）

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，如下公式所示：

 （10）

根据贝叶斯理论，可得：

 （11）

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。

# 2.3 类别不均衡性

## 2.3.1 传统单标签分类中类别不均衡性探讨

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。

## 2.3.2 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。下图展示了常见的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（11），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（12）

，  （11）

 （12）

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。

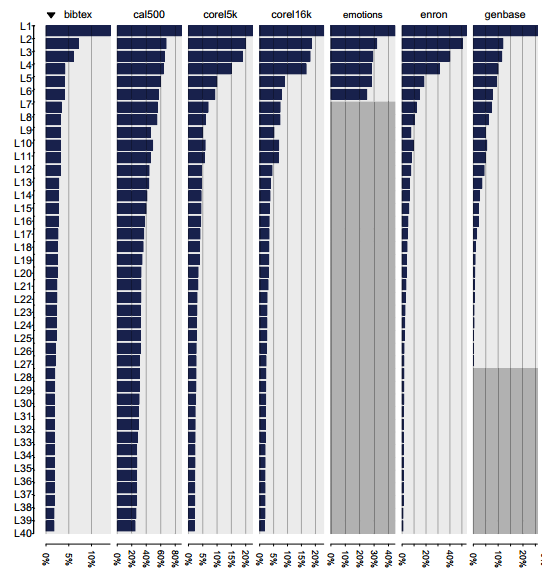


图3 已有多标签数据集统计

## 2.3.3 解决多标签类别不均衡性方法介绍

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：

* 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。

* 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。

* 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。

# 2.4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。

第三章 多标签双重自适应随机采样算法

# 3.1引言

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。

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)。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公共数据集和心血管疾病数据集，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，对应的实验结果及评价。

# 3.2 多标签重采样算法

为了解决多标签学习领域的不均衡问题，目前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（11）(12)，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。

## 3.2.1 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**



## 3.2.2 ML-ROS算法

类似于ML-RUS算法，ML\_ROS算法根据公式（11）（12），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。

## 3.2.3 ML-SMOTE算法介绍

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**



# 3.3 多标签双重自适应采样算法

## 3.3.1 引言

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到的思想，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。

## 3.3.2 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3.2.1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3.2.3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。

## 3.3.3 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（11）（12）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。

## 3.3.4 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，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(13)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。很容易得出，多标签数据

 (13)

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（14），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。

 （14）

## 3.3.5 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**



# 3.4 实验

## 3.4.1 实验数据集

本次实验利用两方面数据集对ML-DARS算法性能进行验证，并与ML-RUS、ML-SMOTE算法进行对比分析。本次实验选取了心血管疾病数据集和4个公共数据集作为实验数据集，其中包括了均衡与不均衡的数据集，具体信息见表4，对这些数据集分别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。

表4 数据集分布指标

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | samples | features | labels | card | MaxIRLbl | MeanIR | meanInstances |
| Cardiovascular |  |  |  |  |  |  |  |
| Chronic |  |  |  |  |  |  |  |
| Emotions |  |  |  |  |  |  |  |
| CAL500 |  |  |  |  |  |  |  |
| medical |  |  |  |  |  |  |  |

## 3.4.2实验设置

## 3.4.3实验结果分析

# 3.5 本章小结

第四章 心血管病预测模型研究

# 4.1 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，从超过4000千万条高血压、糖尿病患者的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而出现异常值，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。

由于原始的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4所示。

图4 模型建立步骤

## 4.2 心血管疾病数据采集

本文中的医学数据针对于高血压、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4，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。

数据来源说明表 4

|  |  |  |  |
| --- | --- | --- | --- |
| 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将所获得数据特征中属于同一大类的疾病进行特征合并，部分信息见表5，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5 疾病归类

|  |  |  |
| --- | --- | --- |
| 疾病代号 | 疾病名称 | 所属类别名称 |
| 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 |

本文的研究疾病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6，这些疾病的人群可重复，因为一位患者可能患有多种疾病。

表6 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 |

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多标签研究对象如表7所示。

表7 研究对象说明

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

## 4.3 心血管疾病数据预处理

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、心肌缺血等疾病，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，利用spark平台和ETL技术处理，具体工作如下：

1. 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。
2. 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8所示。

表8 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 |

1. 由于并发症日期唯一但随访日期不唯一，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近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。
2. 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（13）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(14)所示。

 （13）

为所有样本数据的均值，为所有样本数据的标准差。

 （14）

其中max为样本数据的最大值，min为样本数据的最小值。

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。

# 4.4 心血管疾病数据特征选择

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。

该算法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：

* 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，以心衰为例部分结果如表 9 ，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。

此过程选择出与每个疾病相关性强的特征。

表9 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 |

* 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（15），

 （15）

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。同样以心衰为例，部分结果如表10，年龄、心率等特征权重绝对值大于0.05，保留在心衰的有效特征集合中， 而皮肤病类、代谢综合征等权重过小不作为有效因子。最终合并所有心血管疾病的特征，保留了效果最好的特征集合。

表10 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 |

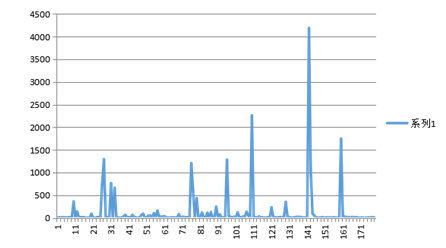
# 4.5心血管疾病预测模型

## 4.5.1 心血管数据集分布

## 4.5.2 建立多标签预测模型

1. 多标签数据集的表现形式为D *= {*(**x**i, Yi) *| 1 ≤* i *≤* m*}* 其中xi代表第i个样本的多标签特征集合，Yi表示该样本的0-1目标向量，向量中的每个值为1代表该样本具有该标签，为0则不具有该标签，预测模型建立步骤分为两个过程，分别如下：

* 首先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，我首先对数据量大的单个标签采用了欠采样策略，多次迭代后多标签数据统计信息中Cardinality为2.24，即数据集中每个标签向量平均有两个标签取值为1，具有多标签分类的意义。但是对于标签向量集合而言依然存在数据不均衡型，如图1表现了多标签向量的数据分布，数据不平衡极为突出，其中大部分多标签向量的样本量在500以下，容易欠拟合，



多标签数据集统计图3

而且也会对模型结果统计评价结果造成不确定性，因此在此基础上我采用了标签向量删除合并策略，

通过前面一系列的处理，应用多标签分类算法评价在当前数据集上的表现。此时约10000多个样本,按70%划分训练集测试集，利用BR-logisticRegression、ClassifierChains、MLKNN等多种多标签算法进行分类。

## 4.6 模型预测能力评估

# 4.7 实验环境介绍

# 4.8 本章小结

第五章 结论与展望

# 5.1 总结

# 5.2 展望

参考文献

[1] Grigorios Tsoumakas, Ioannis Katakis .Multi-Label Classification: An Overview .Dept. of Informatics, Aristotle University of Thessaloniki, 2007.

[2] S Wan , MW Mak , SY Kung . Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017.

[3] Zufferey D, Hofer T, Hennebert J, et al. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. Computers in Biology and Medicine, 2015.

[4] Zhang W, Liu F, Luo L Q, et al. Predicting drug side effects by multi-label learning and ensemble learning . Bmc Bioinformatics, 2015 .

[5] Zhang M L, Zhou Z H. A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 2014.

[6] Boutell M R, Luo J B, Shen X P, et al. Learning multi-label scene classification. Pattern Recognition, 2004.

[7] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification.

Machine Learning, 2011.

[8] Read J. A pruned problem transformation method for multi-label classification[C], 2008.

[9] Tsoumakas G, Katakis I, Vlahavas I. Effective and efficient multilabel classification in domains with large number of labels. ECML/PKDD 2008 Workshop on Mining

Multidimensional Data (MMD’08). (2008) 30–44.

[13] Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition 40 (2007) 2038–2048.

[14]在分类中如何处理训练集中不平衡问题<http://blog.csdn.net/heyongluoyao8/article/details/49408131>

[15] Mining Multi-label Data

[16] Schapire, R.E. Singer, Y.:Boostexter: a boosting-based system for text categorization. Machine Learning 39 (2000) 135–168.

[17] J. He, H. Gu, W. Liu, Imbalanced multi-modal multi-label learning for

subcellular localization prediction of human proteins with both single and

multiple sites, PloS one 7 (6) (2012) 7155.

[18] M.-L. Zhang, Z.-J. Wang, Mimlrbf: {RBF} neural networks for multi-instance

multi-label learning, Neurocomputing 72 (16-18) (2009) 3951–3956, http://

dx.doi.org/10.1016/j.neucom.2009.07.008.

[19] B.-L. Lu, M. Ito, Task decomposition and module combination based on class

relations: a modular neural network for pattern classification, IEEE Trans.

Neural Networks 10 (5) (1999) 1244–1256.

[20] M.A. Tahir, J. Kittler, A. Bouridane, Multilabel classification using

heterogeneous ensemble of multi-label classifiers, Pattern Recogn. Lett. 33

(5) (2012) 513–523, http://dx.doi.org/10.1016/j.patrec.2011.10.019.

1. 摘要中 未突出 工作重点， 未说明具体的工作， 特别是论文思路没有理清 [↑](#footnote-ref-1)