

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4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

April 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等功能，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对象，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，将多标签分类算法应用于心血管疾病预测中，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、多标签分类算法应用于心血管疾病预测时存在的数据严重不均衡性等问题，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，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

Abstract

**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

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。

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。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，BR(binary relevance）是最为常见的一种转化算法。该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。RAk LE（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AdaBoost.MI,Schapire等人提出了用于多标签数据的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。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C4.5算法的多标签决策树，基于支持向量机（support vector machines, SVMs）和神经网络的改进算法等等。

## 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。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