

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

**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。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

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。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，欠采样和过采样并行进行思想，提出了ML-DARS(多标签双重自适应随机采样)算法，均衡数据集分布，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

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

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

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。

## 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：



∆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：



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：

 

 

* 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：





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：



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。



Ranking loss：



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

BR、 HOMER、MLKNN、adboostMH、RAKEL

# 2.3 单标签类别不均衡性

# 2.4 多标签类别不均衡性介绍及方法探讨

2.4.1 多标签类别不均衡性介绍

2.4.2 多标签类别不均衡性评价指标

2.4.3多标签类别不均衡现有方法

类别不均衡问题十分常见，例如，在一个二分类问题中，共有100个样本，其中80个样本属于class 1，其余的20个样本属于class 2，class 1:class2=4:1。在现实中，在欺诈交易识别中，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，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的，只有极少数部分的客户不会再继续享受其服务。

# 2.6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

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.

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)