

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 3](#_Toc382205131)

[1.1 标题2 3](#_Toc382205132)

[1.1.1 标题3 3](#_Toc382205133)

[2 XX（标题1） 3](#_Toc382205134)

[2.1 标题2 3](#_Toc382205135)

[2.1.1 标题3 3](#_Toc382205136)

[3 XXX（标题1） 3](#_Toc382205137)

[3.1 标题2 3](#_Toc382205138)

[3.1.1 标题3 3](#_Toc382205139)

[4 XXXX（标题1） 3](#_Toc382205140)

[4.1 标题2 3](#_Toc382205141)

[4.1.1 标题3 3](#_Toc382205142)

[5 XXXXX（标题1） 3](#_Toc382205143)

[5.1 标题2 3](#_Toc382205144)

[5.1.1 标题3 3](#_Toc382205145)

[6 XXXXXX（标题1） 3](#_Toc382205146)

[6.1 标题2 3](#_Toc382205147)

[6.1.1 标题3 3](#_Toc382205148)

[7 XXXXXXX（标题1） 3](#_Toc382205149)

[7.1 标题2 3](#_Toc382205150)

[7.1.1 标题3 3](#_Toc382205151)

[8 XXXXXXXX（标题1） 3](#_Toc382205152)

[8.1 标题2 3](#_Toc382205153)

[8.1.1 标题3 3](#_Toc382205154)

[9 XXXXXXXXX（标题1） 3](#_Toc382205155)

[9.1 标题2 3](#_Toc382205156)

[9.1.1 标题3 3](#_Toc382205157)

[10 XXXXXXXXXX（标题1） 3](#_Toc382205158)

[10.1 标题2 3](#_Toc382205159)

[10.1.1 标题3 3](#_Toc382205160)

[11 XXXXXXXXXXX（标题1） 3](#_Toc382205161)

[11.1 标题2 3](#_Toc382205162)

[11.1.1 标题3 3](#_Toc382205163)

[12 结论与展望 3](#_Toc382205164)

[12.1 标题2 3](#_Toc382205165)

[12.1.1 标题3 3](#_Toc382205166)

[致 谢 3](#_Toc382205167)

[参考文献 3](#_Toc382205168)

[附 录 3](#_Toc382205169)

[攻读学位期间取得的研究成果 3](#_Toc382205170)

声明

CONTENTS

1 Preface X

1.1 Drip Irrigation Technology X

1.1.1 Drip Irrigation Systems X

1 Preface 11

1.1 Research background and significance 11

1.2 Research status at home and abroad 12

1.2.1 The development of multi-label learning 12

1.2.2 The development of machine learning in the biomedical field 13

1.3 Main research contents 14

1.4 The source of the subject and the structure of the paper 15

2 Multilabel class imbalance and classification methods 17

2.1 Introduction 17

2.2 Multi label learning 18

2.2.1 Formal description of multi label learning 18

2.2.2 The statistical evaluation method of multi label learning 18

2.2.3 Typical processing methods for multi label learning 20

2.3 Class imbalance 24

2.3.1 Class imbalance in traditional single label 24

2.3.2 Class imbalance in multi label 24

2.3.3 The solution of multi label class imbalance 26

2.4 Summary 28

3 Prediction model of cardiovascular disease 29

3.1 The goals and steps of the Model 29

3.2 Data collection of cardiovascular disease 30

3.3 Data preprocessing of cardiovascular disease 32

3.4 Feature selection strategy for cardiovascular disease data 34

3.5 The chapter of summary 36

4 Application of the model of cardiovascular disease 37

4.1 Introduction 37

4.2 Double and adaptive random sampling algorithm in multi label 37

4.2.1 Introduction 37

4.2.2 Multi label resampling algorithm 38

4.2.3 Double and adaptive random sampling algorithm in multi label 42

4.2.4 ML-DARS algorithm description 44

4.3 Experiment 47

4.3.1 Dataset 47

4.3.2 Setting 47

4.3.3 Analysis of the results 48

4.4 The model of cardiovascular disease data 50

4.4.1 Introduction 50

4.4.2 Multi label classification model based on hybrid strategy 51

4.4.3 Experiment 52

4.5 The chapter of Summary 57

5 Conclusion and prospect 58

5.1 Conclusion 58

5.2 Prospect 58

Acknowledgements 59

References 60

Achievements 62

Declarations 63

# 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。



### 多标签学习的统计评价方法

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。

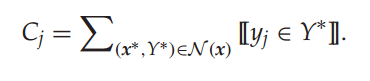


1. 算法适应性方法
2. ML-KNN

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(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。

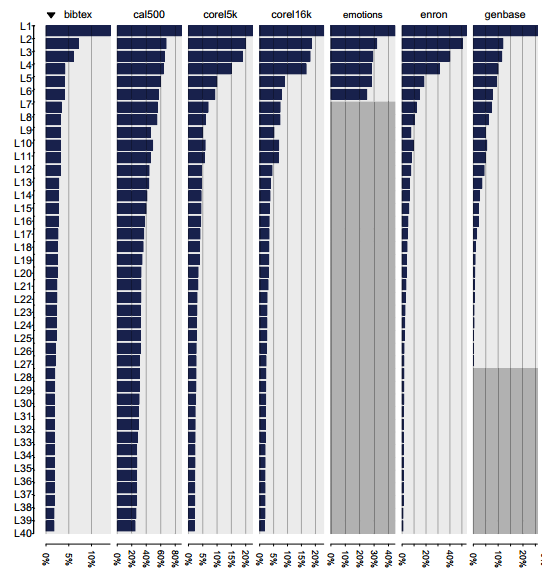


图2-2 已有多标签数据集统计

### 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

# 多标签双重自适应随机采样算法

## 引言

### 标题3

公式按章重新编号：



公式（2-1）说明，…………（公式在正文中的引用）

图题注：

图 2‑1 XXXXXX

# 基于混合策略的心血管疾病预测模型

## 引言

### 标题3

公式按章重新编号：



公式（3-1）说明，…………（公式在正文中的引用）

图题注：

图 3‑1 XXXXXX

# 结论与展望

## 总结

### 标题3

公式按章重新编号：



公式（4-1）说明，…………（公式在正文中的引用）

图题注：

图 4‑1 XXXXXX

# XXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（5-1）说明，…………（公式在正文中的引用）

图题注：

图 5‑1 XXXXXX

# XXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（6-1）说明，…………（公式在正文中的引用）

图题注：

图 6‑1 XXXXXX

# XXXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（7-1）说明，…………（公式在正文中的引用）

图题注：

图 7‑1 XXXXXX

# XXXXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（8-1）说明，…………（公式在正文中的引用）

图题注：

图 8‑1 XXXXXX

# XXXXXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（9-1）说明，…………（公式在正文中的引用）

图题注：

图 9‑1 XXXXXX

# XXXXXXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（10-1）说明，…………（公式在正文中的引用）

图题注：

图 10‑1 XXXXXX

# XXXXXXXXXXX（标题1）

## 标题2

### 标题3

公式按章重新编号：



公式（11-1）说明，…………（公式在正文中的引用）

图题注：

图 11‑1 XXXXXX

# 结论与展望

结论部分着重总结出论文的创新点或新见解及研究展望或建议。

## 标题2

### 标题3

公式按章重新编号：



公式（12-1）说明，…………（公式在正文中的引用）

图题注：

图 12‑1 XXXXXX

# 致 谢

致谢中主要感谢导师和对论文工作有直接贡献和帮助的人士和单位。

一般致谢的内容有：

（一）对指导或协助指导完成论文的导师；

（二）对国家科学基金、资助研究工作的奖学金基金、合同单位、资助或支持的企业、组织或个人；

（三）对协助完成研究工作和提供便利条件的组织或个人；

（四）对在研究工作中提出建议和提供帮助的人；

（五）对给予转载和引用权的资料、图片、文献、研究思想和设想的所有者；

（六）对其他应感谢的组织和个人。

致谢言语应谦虚诚恳，实事求是。字数不超过1000汉字

# 参考文献

（此上两空行不能删除，是为EndNote的参考文献列表所预留）

文后著录的参考文献务必实事求是。论文中引用过的文献必须著录，未引用的文献不得出现。应遵循学术道德规范，避免涉嫌抄袭、剽窃等学术不端行为。

参考文献一般应是作者亲自考察过的对学位论文有参考价值的文献，除特殊情况外，一般不应间接引用。

参考文献应有权威性，要注意引用最新的文献。

参考文献的数量：

硕士学位论文，一般应在30篇以上，其中，期刊文献20篇以上，国外文献10篇以上，均以近5年的文献为主。

对于申请专业学位的学位论文，参考文献的数量可参照执行。

参考文献的著录格式应符合国家标准GB/T 7714-2005《文后参考文献著录规则》。参考文献中每条项目应齐全。

文献中的作者不超过三位时全部列出，超过三位时，一般只列前三位，中文的后面加 “等”字，英文的后面加 “et al”，作者姓名之间用逗号分开。

外国人名一般采用姓在前，名在后的著录法，姓全写且第一个字母大写，名简写成单个大写字母且不加标点，姓和名之间空1格，如：“Metcalf SW”。也可采用名在前，姓在后的著录法，姓全写且第一个字母大写，名简写成单个大写字母且不加标点，名和姓之间空1格，如：“SW Metcalf”。

中文人名的英文表达方式：

简写时，采用姓在前，名在后的著录法，姓全写且第一个字母大写，名简写成单个大写字母且不加标点，如，“钱学森”，简写为“Qian XS ”。

全拼时，名在前，姓在后的著录法，名的第一个字母大写，名连写，名后空1格写姓，姓的第一个字母大写。如，“钱学森”，写为“Xuesen Qian”。

文后参考文献著录格式范例样板，采用五号。

具体要求如下：

A 专著（包括普通图书［M］、论文集和会议录［C］、科技报告［R］、学位论文［D］、标准［S］）

主要责任者．文献题名［文献类型标志］．其他责任者．版本项(第１版不标注) ．出版地：出版者，出版年：引文页码．获取和访问路径．

B 专著中的析出文献

析出文献主要责任者．析出文献题名[文献类型标志]．析出文献其他责任者//专著主要责任者．专著题名：其他题名信息. 版本项(第１版不标注) ．出版地：出版者，出版年：析出文献的起止页码．获取和访问路径．

C连续出版物

主要责任者．题名:其他题名信息［文献类型标志］．年，卷（期）－年，卷（期）.出版地：出版者，出版年．获取和访问路径．

D连续出版物中的析出文献（包括期刊中析出的文献[J]、报纸中析出的文献[N].）

析出文献主要责任者．析出文献题名［文献类型标志］．连续出版物题名：其他题名信息，年，卷（期）：页码．获取和访问路径．

E专利文献

专利发明者/专利申请者或所有者．专利题名: 专利国别,专利号［文献类型标志］.公告日期或公开日期. 获取和访问路径．

F电子文献（包括专著或连续出版物中析出的电子文献）

主要责任者．题名：其他题名信息[文献类型标志/载体类型标志]．出版地：出版者，出版年（更新或修改日期）．获取和访问路径．

表2-2 文献类型和标志代码

|  |  |  |  |
| --- | --- | --- | --- |
| 文献类型 | 标志代码 | 文献类型 | 标志代码 |
| 普通图书 | M | 会议录 | C |
| 汇编 | G | 报纸 | N |
| 期刊 | J | 学位论文 | D |
| 报告 | R | 标准 | S |
| 专利 | P | 数据库 | DB |
| 计算机程序 | CP | 电子公告 | EB |

表2-3 电子文献载体和标志代码

|  |  |  |  |
| --- | --- | --- | --- |
| 载体类型 | 标志代码 | 载体类型 | 标志代码 |
| 磁带（magnetic tape） | MT | 磁盘（disk） | DK |
| 光盘（CD-ROM） | CD | 联机网络（online） | OL |

样例：

1. 刘国钧,郑如斯．中国书的故事［M］．北京：中国青年出版社，1979：110-115．
2. 昂温 G． 外国出版史[M].陈生铮译．北京：中国书籍出版社，1988．
3. 辛希孟．信息技术与信息服务国际研讨会论文集：A集［C］．北京:中国社会科学出版社,1979．
4. 冯西桥．核反应堆压力容器的LBB分析［R］．北京：核能技术设计研究院，1997．
5. 张和生．地质力学系统理论［D］．太原：太原理工大学，1998．
6. 全国文献工作标准化技术委员会第七分委员会．GB/T 5795-1986．中国标准书号［S］．北京:中国标准出版社，1986．
7. 罗云．安全科学理论体系的发展及趋势探讨[M]//白春华，何学秋，吴宗之．21世纪安全科学与技术的发展趋势．北京：科学出版社，2000：1-5．
8. 钟文发．非线性规划在可燃毒物配置中的应用[C]//赵玮．运筹学的理论与应用：中国运筹学会第五届大会论文集．西安：西安电子科技大学出版社，1996：468－471．
9. 高义民，张凤华，邢建东等．颗粒增强不锈钢基复合材料冲蚀磨损性能研究[J]． 西安交通大学学报，2001，35(7)：727-730．
10. Papworth A, Fox P, Zeng GT, et al. Ability of aluminum alloy to wet alumina fibres by addition of bismuth[J]. Mater Sci & Technol,1999,15(4):419-428.
11. 丁文祥．数字革命与竞争国际化[N]．中国青年报，2000－11－20(15)．
12. 姜锡洲．3一种温热外敷药制备方案：中国，881056078［P］．1989-07-26．
13. Koseki A,Momose H,Kawahito M,et al Complier:US,828402［P/OL］2002-05-25 [2002-05-28].http://FF&p.
14. Online Computer Library Center, Inc. History of OCLC[EB/OL].[2000-01-08]. http://www. clc.org/ about/history/default.htm.
15. 江向东．互联网环境下的信息处理与图书管理系统解决方案[J/OL]．情报学报,1999,18(2):4[2000-01-18].http://www.chinainfo.gov.cn/periodical/qbxb．
16. Scitor C. Project scheduler[CP/DK].Sunnyvale,Calif.:Scitor Corp, 1983.
17. Metcalf SW. The Tort Hall air emission study[C/OL]//The International Congress on Hazardous Waste, Marquis Hotel, Atlanta,Georgia,June 5-8,1995: impact on human and ecological health[1998-09-22]. <http://atsdrl>.atsdr.cdc.gov:8080/cong95. html.

参考文献里面标点符号：英文文献用半角,中文文献用全角。

# 附 录

附录编号依次编为附录A，附录B。附录标题各占一行，按一级标题编排。每一个附录一般应另起一页编排，如果有多个较短的附录，也可接排。附录中的图表公式另行编排序号，与正文分开，编号前加“附录A-”字样。

本部分内容非强制性要求，如果论文中没有附录，可以省略《附录》。

# 攻读学位期间取得的研究成果

1）已发表或已录用的学术论文、已出版的专著/译著、已获授权的专利按参考文献格式列出。

2）科研获奖，列出格式为：

获奖人（排名情况）．项目名称．奖项名称及等级，发奖机构，获奖时间．

3）与学位论文相关的其它成果参照参考文献格式列出。

4）全部研究成果连续编号编排。

样例：

1. Wei ZY, Tang YP, Zhao WH, et al. Rapid development technique for drip irrigation emitters[J]. RP Journal,UK., 2003, 9(2):104~110 (SCI: 672CZ; EI: 03187452127).
2. 魏正英,唐一平,卢秉恒.滴灌管内嵌管状滴头的快速制造方法研究[J].农业工程学报, 2001,17(2):55~58 (EI:01226526279,01416684777).

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)