学校编码：10384　　分类号____密级____
学号:23020141153151UDC____

厦门大学

硕士学位论文

# 基于标签相关性和三层 BP 神经网络的多标签分类算法研究

# Research on Multi -label Classification Algorithm Based on Label Correlation and Three - layer BP Neural Network

廖丽芳

指导教师姓名：吴梅红副教授

专 业 名 称：计算机科学与技术

论文提交日期：2017 年　月

论文答辩日期：2017 年　月

答辩委员会主席：

评　阅　人：

年　月

# 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　月　日

# 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　年　月　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

# 摘要

现如今，多标签分类已然是各大机器学习相关会议上的热门讨论话题。在传统的分类问题中，一个样本只能对应一个标签，但是在多标签分类问题中一个样本却往往对应一个标签集合，这个标签集合中通常含有2个或者2个以上的标签，因此传统的方法已无法满足多标签分类的需求。多标签分类问题在现实生活中经常出现，近十几年来，国内外学者对这个问题经过不断地探索产生了一些相关成果，发表了各式各样的多标签分类算法。其中一种利用三层BP神经网络的多标签分类算法表现出较好的性能，且应用领域广，可以应用到各种不同的分类数据库中，但是这种算法往往都存在一些不足，比如算法性能低下、算法运行时间较长等。究其原因，很大部分是由于以下两个部分：一方面有些算法未考虑标签之间的相关性，另一方面有些算法并不能完整地或者正确地描述标签之间的相关性。因此针对这种算法还是存在许多可以研究的方面。

本文中我们首先从多标签分类问题目前的研究现状出发，然后进行了相关问题的描述，并讨论了解决该问题面临的挑战。总结了目前多标签分类算法的性能衡量指标和评估方法。详细地回顾了三层BP神经网络的基础知识，包括网络结构、学习过程、训练流程、网络中一些参数的选取方法，还介绍了对输入数据进行预处理的一些方法。接着本文深入地总结了目前利用三层BP神经网络进行多标签分类的四种算法，并采用实验证明了此类算法中考虑标签之间关系的算法有一定的优势。最后在前人的基础上，重新提出了一种利用标签相关性和三层BP神经网络的多标签分类算法。在该算法中充分考虑了标签集合中相关标签与不相关标签间的关系、相关标签间的关系和不相关标签间的关系。并通过对比试验证明该算法相比本文中其他算法表现更好。

关键词：多标签分类；三层BP神经网络；标签相关性；

# ABSTRACT

Nowadays, multi-label classification is already a hot topic for discussion at all major machine learning conferences. In the traditional classification problem, a sample can only correspond to a label. But in a multi-label classification problem, a sample often corresponds to a set of labels, which usually contain two or more labels. So the traditional method has been unable to meet the needs of multi-label classification. Multi-label classification problems often occur in real life. Over the past decade, researchers at home and abroad have explored this problem and produced some related results, and published a variety of multi-label classification algorithms. One of the Multi-label classification algorithms using three-layer BP neural network shows good performance, and the application field is wide and can be applied to a variety of different classification databases. But this algorithm often has some shortcomings, such as algorithm performance Low, the algorithm runs longer. The reason is largely due to the following two parts: on the one hand, some algorithms do not take into account the correlation between labels, on the other hand some of the algorithms can not completely or correctly describe the correlation between the labels. So there are still many aspects of this algorithm can be studied.

In this paper, we start from the current research status of multi-label classification, and then describe the relevant issues, and discuss the challenges to solve the problem. This paper summarizes the performance measurement and evaluation methods of the current multi - label classification algorithm. The basic knowledge of the three - layer BP neural network is reviewed in detail, including the network structure, the learning process, the training process, the selection of some parameters in the network, and some methods of preprocessing the input data. Then, this paper summarizes four algorithms which use three-layer BP neural network for multi-label classification, and prove that there are some advantages in the algorithm to consider the relationship between labels. Finally, a multi-label classification algorithm based on label correlation and three-layer BP neural network is proposed on the basis

of predecessors. In this algorithm, the relationship between the relevant label and the irrelevant label in the label set, the relationship between the relevant labels and the relation between the irrelevant labels are fully considered. And the experimental results show that the algorithm show better performance compared to other algorithms in this paper.

Key Words: multi-label classification; three-layer BP neural network; label correlation;

# 目录

# Contents

# 第一章 绪论

## 1.1 研究背景与意义

去年 3 月，谷歌开发的智能程序与世界冠军进行了人机大战，最终这场比赛以智能程序大比分领先落幕。这场比赛的胜利让人工智能一时间进入了公众的视野。现今，机器学习已成为人工智能的重要组成部分，包含计算机技术、工程技术、数理统计，横跨政治、医学、金融、地质学等多个领域，并能用来支持我们日常生活应用，例如人体轮廓识别、垃圾邮件分类、手写体识别、产品推荐、顾客分析，使我们的生活变得更加的智能化。利用机器学习，我们可以从无序的数据中找出规律，然后利用该规律获取更多潜在的有效信息。机器学习的主要任务就是分类，属于监督学习，分类就是确定对象属于哪个预定义的标签（即目标类），例如：利用电子邮件的标题和内容区分邮件的有效性，利用医院设备采集的影像判断肿瘤是恶性还是良性的，利用语音特征信号进行音乐的分类。其中如果预定义标签集合中只有一个或者两个标签，则该类问题称为二值分类。如果预定义标签集合中含有标签达到两个以上，则该类问题称为多类分类。以上两种都属于单标签分类问题，即一个样本只能与一个标签关联。但是在现实生活中，由于数据的多样性，可能存在一个样本同时与超过两个类别标签关联的特殊情况，我们称之为多标签分类问题。

多标签分类问题涉及的主要领域很多，包含图片场景数据、文本数据以及生物医学数据等的分析。例如：在场景分类中，一张图片可能包含多个场景，关于自然场景的图片中，这张图片可能同时包含“树木”、“天空”、“大海”这些标签。在文本分类中，一篇文档可能含有多个预定义的主题，关于南非足球杯比赛胜利的新闻报道中，该篇文章的主题可能包含‘非洲’、‘体育’这些标签。在邮件主题分类中，每封邮件可能含有多个特点，如“私人的”、“情感为喜悦”、“主题为祝贺”。在生物信息学中，每个基因可能含有多种功能，如“新陈代谢功能”、“转录功能”、“蛋白质合成功能”；由此可知，随着世界的快速发展，数据量越来越庞大，数据本身越来越复杂，多标签分类问题更加常见。因此对多标签分类算法

的研究具有深远的现实意义和价值。但是解决多标签分类问题的过程中却会遇到许多的挑战，例如多标签分类需要面临巨大的输出空间，即可能的标签集合的数量达到指数级。多标签分类的过程对应于特征空间到标签子集空间的映射过程，其中标签子集的空间等价于标签的幂集。因此，当存在大量甚至中等数量的标签时，标签的幂集数目可达到百万级。

人工神经网络（ANN）是一种模拟人体大脑活动的智能信息处理系统，在生物科学、金融分析、医疗诊断等方面的应用取得了不错的成果。从数学的方面来讲，人工神经网络可以认为是一种特殊的运算，给定相应的输入信息，让信息在神经元之间进行传递，最终得到对应的输出信息。人工神经网络能够快速且准确地进行自我学习且泛化能力突出，在处理一些复杂的问题上表现出一定的优势。

经过不断研究创新，D.Rumelhart 等人在 1985 年提出 BP（Back—Propagation）神经网络。在 BP 神经网络中，输入信息依次往前传递，输出值与期望输出间的误差则往相反的方向传递。BP 神经网络的结构简单，易于理解，自我调整能力较好且操作起来比较容易，具有自组织、自适应性以及较强的鲁棒性。BP 神经网络算法通过有效学习，可以完成其它方法无法完成的特定任务，因此被大量使用于模式分类、系统仿真、图像识别等方面。许多学者开始将 BP 神经网络用于解决分类问题，并取得了不错的成果[32],[33],[34],[35]；其中 T. F. Burks 使用 160 张不同纹理图像为数据样本，使用 BP 神经网络进行图像分类，并取得了高达 96.7% 的正确率[32]。该学者还使用著名的地球卫星遥感图像数据集，探讨了不同 BP 神经网络结构对于图像分类的影响[33]。Justin D. Paola 等人使用图森、美国亚利桑那州、奥克兰、加利福尼亚四个城市的土地图片为数据集，利用 BP 神经网络进行图像分类，并与最大似然方法比较[34]。P. Wilding 等人使用人体的医疗数据，利用 BP 神经网络进行乳腺癌和卵巢癌的诊断[35]。三层 BP 神经网络已经被证明：当隐含层神经元个数达到一定程度，它就具有实现非线性系统的能力[55],[56]。

通过以上内容可知，将三层 BP 神经网络用于多标签分类的探索具有一定的现实意义和价值。