

毕业设计（论文）开题报告

题目： 基于python的奥运会比赛的数据挖掘

专业 计算机科学与技术（高本贯通）

班级 207041G1

学号 207041G113

姓名 瞿泽炜

指导教师 王浩

2022年1月5日

# 本课题的目的和意义、国内外研究现状、水平和发展趋势

## 课题的目的和意义

随着去年落下帷幕的东京奥运会以及今年拉开序幕的北京冬奥会，人们的奥运热潮升温。因此，针对热度不减的奥运热潮，我们首先将对奥运会比赛背后所蕴含的各项数据进行深度的挖掘和处理，形成可视化的数据分析，从而揭露其中的规律和变化，让人们能够更加清晰地了解奥运比赛，通过数据读懂奥运比赛，对数据产生敏感以及安全性的保护。其次，以中国为例，通过可视化的奥运历史数据，发现数据中的关联规则和趋势走向，建立基于时间序列的奖牌回归模型以生成下届夏季奥运会比赛中中国金牌、奖牌数的预测性数据分析，最后实现对中国下届奥运比赛成绩的科学性预测。

## 国内外研究现状与水平

像奥运会这样的大型赛事，在近年来受到了国内外许多专家的关注，他们对于奥运比赛的数据产生了浓厚的兴趣，不单单是赛场内的，赛场外的也同样如此。通过建立数据挖掘模型，对金牌榜和奖牌榜进行预测和分析。

自1896年现代奥运会举办以来,奥运会奖牌的数量是世界各国关注的热点,其原因之一,就在于奖牌数量反应了参赛国的体育竞技实力。但以往关于体育运动的研究,大多是从体育专业的角度出发,较少有将数据挖掘理论应用于体育或者奥运会的研究,因而对它的研究是缺乏数字依据的。中国奥委会是在1979年国际奥委会合法席位得到恢复后, 于1980年首次出席美国普莱西德湖第13届冬季奥运会, 1984年首次派代表团参加了美国洛杉矶第23届夏季奥运会, 由此揭开了新中国征战奥运会的历史, 也开始了中国体育科学研究领域奥林匹克研究的历史。近年来,奥运会奖牌预测不断升温,并在短时期内就有不同的学科领域提出了多种预测模型方法,其理论方法和预测模型在迅速地发展和完善。在国外迅速发展的背景下,国内的少数学者近两年来也极为关注预测方法的进展,并进行了尝试性实践。

当前，随着社会上博彩行业的发展，一些体育和娱乐赛事预测比赛结果的行为不断泛滥，我们暂且不论这种行业是否违法，至少这个行业的确带动了体育相关产业的发展，更重要的是带动数据挖掘技术在体育数据分析领域的发展。比赛预测的主要目的和作用包括两大方面：验证和发现对成绩有意义的影响因素；对赛事的结果进行预测。比赛预测不同于比赛现场数据挖掘技术，因为他们所利用的数据库是完全不同的，比赛预测也就是在比赛之前，利用以往的比赛或训练数据作为基础，从而判断获胜的概率，能解决参赛群体的多样性、多变性、影响因素的复杂性和预测的困难性等问题。该技术已经在彩票等体育产业领域广泛应用，但是其预测的准确性和稳定性很难得到控制，仅能作为一种参考。

存在的问题：

（一）理论研究不足

目前数据挖掘理论相对来讲还比较抽象，特别是体育领域缺乏一些计算机专业人才，从而导致该领域的研究不够深入。通过纵向对比，我们发现其他学科领域数据挖掘技术的研究相对较多，这得益于他们对数据挖掘应用原理与技术的研究相对比较深入。国内外虽有很多学者已投入该领域研究中，但至今未出现有影响力的挖掘模型，同时成果的数量与内容都反映出对体育数据挖掘的理解还停留在较浅层面。

（二）结果有效性与确定性

数据挖掘处理对象是海量样本，而且数据搜集、归类、整理、分析是一系列繁琐的过程，一旦一个环节出现错误将会影响预测效果，并且数据挖掘必须建立在海量样本的基础之上，只有提高数据分析量，才能提高结果的有效性与科学性。我们从大部分研究论文中发现，实际分析案例数据量小，且有些还是模拟数据，并非真实数据对小样本数据，寻找的规律或结果可能不具有统计显著意义，不能反映体育变化规律或结果的有效性。

（三）研究的手段和方法缺乏灵活性

目前数据挖掘技术在体育领域的运用还基本是依靠计算机领域的传统研究方法和手段，当研究者需要哪一部分数据和参考值，就会临时搜集部分数据库，根本不对数据库的信息进行有效分析，有些甚至不懂得哪些是重要因子，哪些数据具有对参考值比较高的响应率等，数据挖掘的方法和手段只是模仿和照搬。但事实上数据挖掘技术，是依靠多方面的数据资料库，来分析体育领域存在的问题和优势，进一步推进体育领域与计算机领域的交叉和融合，有利于产生更合理、更科学的研究手段和方法，对促进数据挖掘技术、体育技能的发展和普及具有重要意义。

## 发展趋势

祝玮东在《当代体育研究中数据挖掘技术的应用》中认为，当今计算机的计算能力正日益提高，各领域业务复杂性也不断提升，数据的类型越来越多、越来越复杂。数据挖掘的服务内容是为体育科研人员提供有效的数据，如何提高服务质量，如何将有价值的知识资源提供给有需求的用户是我们研究的出发点。

我国体育领域积累并不断扩充着大量的数据，我们的体育科研人员要充分重视这些数据的整理，建立有效的数据仓库和数据集，并运用数据挖掘进行分析，创造新的价值，这是体育科研人员今后运用数据挖掘技术的重要任务，这也为奥运比赛的全面发展提供了强有力的支持，通过数据挖掘让奥运会这样的全球性赛事平台在全世界范围内起到进一步的发展和引领作用。

# 文献查阅、调研情况

## 文献综述

自诞生之日起,奥运会比赛中的各种数据成为了各界专业人员研究和探索的原因，而历届奥运会奖牌就此作为了专门的、系统的奖牌预测研究。艺术品被世人珍藏,凝结着奥林匹克精神与各民族风情特色。因此,对奥运会奖牌的预测也吸引了各界学者。

第二次世界大战结束后,一些社会学家和经济学家从社会学与经济学角度预测分析了影响国家获奥运会的奖牌数。较早关于这方面的研究成果有Ball(1972)、Grimes AR等人(1974)和LevineN(1974)。一个有趣的现象是,近30年后关于这方面的理论研究才重新被人们所关注。Shughart等人(1993)首次重新利用过渡型经济方法预测分析了世界各国在奥运会上取得的成绩。Johnson和Ali(2000)与Bernard和Busse(2000)重新将人们的目光聚集到该研究上,人们才开始对奥运会进行专门的、系统的奖牌预测研究。

李响（2018）从奥运会历史数据中发现时序关联,,建立预测模型具有深远意义。以法

国历届奥运会奖牌数作为实验数据,在R语言下实现了时间序列模型的建立、检验并预测出法国下届奥赛的奖牌数。从中也可以看出时间序列分析应用于奥运会成绩的挖掘具有一定的科学性，对于奥运比赛的数据挖掘应该全面考虑相关因素，采用多种方法进行挖掘和分析。

杨华磊、周晓波（2012）以数据挖掘的视角研究了奥运会数据背后所凸显的唯象法则，分析奥运奖牌数的分布情况以及其中蕴含的普适性法则。

杨秉龙（2010）总结了奥运会上我国运动员所获奖牌的历史经验, 深入探讨所获奥运会奖牌的内在规律, 对我国今后制定和实施奥运战略具有重要的意义与价值。采用文献资料法、数理统计法和逻辑分析法, 对中国获得的奥运会金牌、奖牌状况进行统计分析, 归纳总结获得金牌、奖牌的特征, 并对特征进行分析研究。

刘龙祥（2018）利用自回归分析和ARIMA（1,1）模型进行了预测分析，对2020年各国奥运会金牌数进行预测,最终对2020年奥运会金牌榜前十名进行了预测。研究了各国在奥运中获得的金牌数的时序关系，采用自回归滑动平均模型，以1996年—2008年的奥运会成绩作为训练集，以2012和2016年作为预测集。

根据所描述的研究文献来看，对于奥运会比赛的数据挖掘大多都以奥运奖牌数、奥运历史数据的挖掘为主，主要分析奖牌数分布情况的特征和关联因素。普遍都采用了时间序列预测分析和多元线性回归预测分析两种不同的方法对奥运会比赛的奖牌数进行了预测。

## 调研情况

从国际奥委会官网和中国奥委会官网上查找关于奥运会比赛的介绍、历史背景、项目组成、参赛人数、奖牌分布状况等相关的信息数据以及中国队在奥运比赛上的表现和获奖情况。进入Kaggle网站，对奥运比赛的相关数据进行更加深入的调查和研究。

与此同时，通过参考相关网站和国内外的文献资料对基于python的数据挖掘技术进行了研究，了解如何用python进行数据挖掘，有何注意事项。初步了解和学习了可视化数据的分析方法以及时间序列预测法的基本步骤和特征。

目前，已对研究所需的各种数据和概念有了一定的了解，研究方向和思路初步定型，但具体方法的使用仍没有得到完善，对于可视化数据的呈现方式以及如何通过python实现有些许理解不足有待日后进一步研究和优化，而对于时间序列预测法中模型的建立以及深入计算仍有欠缺之处。

# 本课题的基本内容、重点、难点

## 本课题的基本内容

1. 通过python爬取国际奥委会官网以及其他各大网站关于奥运比赛的相关数据，导入Kaggle网站上的数据。数据包括从1896年奥运会到2020年奥运会（夏季奥运会和冬季奥运会）每届奥运比赛参赛队伍的数量、参赛人数、男女人数的比例以及比赛项目的数量。
2. 将所爬取和导入的数据逐个实现数据可视化呈现。
3. 对奥运比赛的各项历史数据进行分析，认识奥运比赛的发展历程——根据可视化数据进行深入分析,计算出奥运比赛参赛队伍数量、参赛人数、比赛项目数量的增长率以及男女人数的比例，从而通过增长率预估下届奥运会比赛的参赛队伍数量、参赛人数、比赛项目数量。
4. 引起人们对数据安全的重要性，从而采取反爬取技术的数据保护手段。
5. 针对中国代表队在所参加的历届夏季奥运比赛的奖牌总数，爬取和导入所需数据。
6. 将数据进行可视化呈现。
7. 分析奖牌数的变化趋势。
8. 进行数据异常性和有效性检测等构建预测模型前相关的准备工作。
9. 建立时间序列的奖牌回归模型。
10. 通过模型对中国2024奥运会的奖牌数进行预测。

## 本课题的重点

数据的选取和处理、通过python实现所收集数据进行可视化的分析、时间序列的分析、选取何种时间序列预测模型、如何成功构建模型并进行预测。

## 本课题的难点

各个网站数据爬虫的解析和提取并实现可视化、不同可视化图表的绘制方法、各种数据的计算、时序关系的分析方法、各种函数的使用、实现预测前做的各项准备工作。

# 解决问题的方法、手段、措施等

## 拟采取的方法和技术

方法：

文献研究法：上网搜索相关资料，查阅与课题相关的文献，对本课题涉及到的知识和数据有一定了解之后，形成了较为清晰的研究思路。

功能分析法：有了与课题相关的知识储备之后，对课题的设计过程进行科学性和逻辑性规划，确定设计的环境和工具，划分软件中各个模块的功能。根据需求分析各个模块所能实现的具体功能以及整个软件所要实现的功能，整合部分模块的功能来达到整体实现的效果，将不同模块之间的功能衔接和转换得到进一步完善。

实验法：有了所需模块的功能分析后，用python语言在设计工具中进行代码的编写，不断调试和修改，实现各个模块的功能以符合设计的目的。

技术：

爬虫技术：爬虫是模拟用户在浏览器或者App应用上的操作，把操作的过程、实现自动化的程序。它的工作原理：发起请求→获取响应内容→解析内容→保存数据。使用 requests 库抓取网站上的数据。

数据分析技术：使用NumPy、matplotlib以及pandas等模块实现数据的可视化和计算。

反爬虫技术：即为使用任何技术及手段，阻止被人抓取自己网站信息的一种方法，关键在于批量和减少阻止过程中的误伤。一般网站从三个方面反爬虫：1.用户请求的Headers，2.用户行为，3.网站目录和数据加载方式。

时间序列预测技术：划分训练集和测试集数据，依靠过去和现在的数据，分析两者之间的关系，然后利用得到的这个关系去预测未来的数据。

## 选择的工具

选择Jupyter Notebook应用程序进行设计，使用python语言进行编写。

Jupyter Notebook可以直接在代码旁写出叙述性文档，而不是另外编写单独的文档。也就是它可以能将代码、文档等这一切集中到一处，让用户一目了然。适合做数据分析这些需要可视化的操作。而python语言作为Jupyter Notebook中所涉及到的三种科学运算语言之一，简单易用，使用成本低，功能强大。

## 工作进度安排

2021年12月11日~2021年12月25日确定指导老师、选题

2021年12月25日~2022年1月4日搜集相关资料、准备开题报告

2022年1月5日~2022年1月16日完成开题报告以及外文翻译

2022年1月16日~2022年2月1日确定设计方案和论文撰写提纲

2022年2月1日~2022年4月1日完成论文初稿

2022年4月1日~2022年5月1日完成论文修改并提交成稿

# 论文提纲

摘要

关键词

1. 绪论

1.1选题背景

1.2研究目的及意义

1.3国内外研究现状

1.4研究内容

第二章理论基础

2.1使用环境

2.2相关概念

2.3数据来源

2.4研究方法

第三章奥运会比赛历史数据挖掘

3.1数据准备

3.2数据预处理

3.3数据可视化

3.3分析和计算

3.4数据保护

3.5小结

第四章中国2024奥运会比赛奖牌数预测

4.1数据准备

4.2数据预处理

4.3数据可视化

4.4数据分析

4.5构建预测模型并预测

4.6小结

第五章结束语

参考文献

致谢

# 主要参考文献

1. 刘龙祥.[基于数据挖掘模型对2020年奥运金牌数的预测研究](https://kns.cnki.net/KNS8/Detail?sfield=fn&QueryID=2&CurRec=4&recid=&FileName=CXJL201805132&DbName=CJFDLAST2018&DbCode=CJFD&yx=&pr=&URLID=)[J].信息记录材料，2018，19(05)：203~205.
2. 丁尉哲.基于综合国力的奥运奖牌数据挖掘模型[J].信息记录材料,2018,19(03)：231~233.
3. 朱坚.[男子1500米自由泳竞技能力时间序列特征数据挖掘研究](https://kns.cnki.net/KNS8/Detail?sfield=fn&QueryID=2&CurRec=7&FileName=1016153862.nh&DbName=CDFDLAST2018&DbCode=CDFD)[D].北京体育大学.2016
4. 杨华磊，周晓波.奥运奖牌数据背后所凸显的唯象法则——基于数据挖掘视角的探究[J].2012,32(10)：3~11.
5. [苏旸](https://kns.cnki.net/KNS8/Detail?sdb=CJFD&sfield=%e4%bd%9c%e8%80%85&skey=%e8%8b%8f%e6%97%b8&scode=000005580221&acode=000005580221).[第29届奥运会奖牌分布的特征研究](https://kns.cnki.net/KNS8/Detail?sfield=fn&QueryID=2&CurRec=11&recid=&FileName=BJTD200906034&DbName=CJFD2009&DbCode=CJFD&yx=&pr=&URLID=)[J].北京体育大学学报,2009,32(06)：105~108.
6. 李响.[基于时间序列分析的奥运会奖牌数预测](https://kns.cnki.net/KNS8/Detail?sfield=fn&QueryID=2&CurRec=3&recid=&FileName=JSSG201803024&DbName=CJFDLAST2018&DbCode=CJFD&yx=&pr=&URLID=)[J].计算机与数字工程,2018,46(03)：533~536,545.
7. 管河山，邹清明，罗智超.时间序列平稳性分类识别研究[J].统计与信息论坛,2016，31(04)：3~8.
8. 王娜.时间序列建模、预报的原理[J].吉林工程技术师范学院学报，2012，23(3)：78~80.
9. 刘罗曼.时间序列平稳性检验[J].沈阳师范大学学报（自然科学版）,2010,28(3)：357~359.
10. MandrikovaOksana, FetisovaNadezhda, Polozov Yuriy. [Hybrid Model for Time Series of Complex Structure withARIMAComponents[J]](https://kns.cnki.net/KNS8/Detail/RedirectScholar?flag=TitleLink&tablename=GARJ2021_1&filename=SJMDABB02CE34DE3818EF8FEBFE384A03B77).Mathematics,2021,9(10):1122.
11. Tingting Cheng, Functional. 2019,38(06)：636~659.