Hans Journal of Data Mining 数据挖掘, 2022, 12(4), 310-319
Published Online October 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.124030

Hans汉斯

# 基于半监督聚类的NBA季后赛第一轮预测

刘　凯[1]，李卓雅[2]，李新民[1*]

[1]青岛大学数学与统计学院，山东 青岛
[2]青岛海尔智能家电科技有限公司，山东 青岛

收稿日期：2022年8月20日；录用日期：2022年9月20日；发布日期：2022年9月28日

## 摘　要

**体育赛事的兴起使得大量的数据被纪录下来，体育统计随之发展起来。在众多体育赛事中，NBA是其中一个影响力较大的体育联盟，在NBA数据的分析中季后赛预测是一个重要的方面。NBA季后赛分为四个阶段，将预测分为多阶段分析建模并进行预测有着现实意义，本文旨在研究季后赛第一轮的预测问题。季后赛的预测实际是一个二分类问题，本文通过整理当前赛季常规赛的比赛统计数据、教练的历史执教数据和球员当前赛季的RPM值，进而从球队、教练、球员三个方面给出球队实力的评价值，并在此基础上建立有勿连约束和必连约束的半监督聚类模型，最后根据历史统计数据给出已分好类的标签，预测结果表明半监督聚类在NBA季后赛第一轮的预测中有着较好的预测效果和很强的适用性。**

## 关键词

**季后赛预测，半监督聚类，球队实力评价**

# Prediction of the First Round of NBA Playoffs Based on Semi-Supervised Clustering

**Kai Liu[1], Zhuoya Li[2], Xinmin Li[1*]**

[1]School of Mathematics and Statistics, Qingdao University, Qingdao Shandong
[2]Qingdao Haier Intelligent Home Appliance Technology Co. Ltd., Qingdao Shandong

Received: Aug. 20th, 2022; accepted: Sep. 20th, 2022; published: Sep. 28th, 2022

## Abstract

**With the rise of sports events, a large number of data have been recorded, and sports statistics have developed accordingly. Among many sports events, the NBA is one of the most influential**

*通讯作者。

文章引用: 刘凯, 李卓雅, 李新民. 基于半监督聚类的 NBA 季后赛第一轮预测[J]. 数据挖掘, 2022, 12(4): 310-319.
DOI: 10.12677/hjdm.2022.124030

**sports leagues, the prediction of the playoffs is an important topic in the study of NBA data. The NBA playoffs are divided into four stages. It is of great practical significance to divide the prediction into multiple stages. This paper aims to study the prediction of the first round of the playoffs. The prediction of the playoffs is actually a two-category problem. By arranging the game statistics of the regular season, the historical coaching data of the coach and the RPM value of the players in the current season, we provide the evaluation value of the team's strength from the three aspects of the team, the coach and the players. On the basis of establishing data, a semi-supervised model with must-link and cannot-link constraints is established. Finally, according to the historical statistical data, the well-classified labels are given. The prediction results show that semi-supervised clustering has a good prediction effect and strong applicability in the prediction of the first round of the NBA playoffs.**

## Keywords

**Playoff Forecast, Semi-Supervised Clustering, Team Strength Evaluation**



## 1. 引言

体育赛事在人们的日常生活中有着越来越重要的影响，伴随着体育赛事的兴起，大量的体育数据被统计记录下来，如何更好的利用这些体育数据也成为一个重要的研究方向，体育统计随之发展起来。在体育统计领域，机器学习和数据挖掘的应用也越来越广泛[1]。用过去和现在的数据去分析球员的能力一直有很多的研究，特别是在篮球领域[2] [3]。利用这些数据通过多种方法去进行预测和分析比赛的结果也变得越来越普遍[4]。在篮球领域，NBA (National Basketball Association)是其中发展较为职业化，国际影响力较大，品牌影响力较强的一个联盟，对于 NBA 的统计分析有着很强的需求。NBA 非常注重体育赛事数据的收集，也非常注重这些数据的应用。NBA 的数据有着很多方面的应用，例如球员的价值分析、常规赛 MVP 预测和季后赛预测等等。在众多的分析中，NBA 季后赛的预测是一个重要的方面，NBA 季后赛是球队一个赛季努力的目标，进入季后赛的球队可以获得球场的门票收入和更大的知名度从而获得巨大的商业价值。

NBA 季后赛是 NBA 常规赛成绩排名靠前的球队进行综合角逐的比赛。具体来说 NBA 共 30 只球队，东西部各 15 只，东西部分别排名前 8 的球队进入季后赛。NBA 的季后赛通常分为四个阶段，分别为季后赛第一轮的比赛、分部准决赛，分部决赛以及 NBA 总决赛。对于一场系列赛来说影响球队胜利的因素有很多，包括球员、教练、主客场以及球队的健康情况等。教练是球队的指挥者，球员是球场上的竞争者，王明新通过分析 2015-16 赛季总决赛的制胜因素认为教练是一轮系列赛制胜的关键，球队的攻防战术也是胜负的核心影响因素[5]。勒勇等通过分析 2005-06 赛季总决赛的数据，认为总决赛的制胜因素为投篮命中率、三分命中率、篮板球、助攻、盖帽等技战术水平发挥水平、球星的数量和质量以及主教练的总决赛经验和心理调节能力[6]。

NBA 的季后赛预测实际是一个二分类的问题，针对分类问题常见的学习算法有逻辑回归、朴素贝叶斯、支持向量机、随机森林等监督学习算法和半监督聚类、半监督 SVM 等半监督学习算法。邱胜等将 2004 到 2006 三个赛季的常规赛数据从球员、球队、主客场三个大的方面进行整理后建立逻辑回归和贝

叶斯模型进行季后赛的预测，其认为主场优势对于胜负有着较大的影响，其提出并采用了一个新的因子分析计算方法评估球员的能力优于 Oliver 的公式[7] [8]。Galsanbadam 等通过支持向量机回归、多项式回归和决策树回归的方法评估了个人球员的表现对球队胜率的影响[9]。Cheng 等通过建立 NBA 极大熵模型来预测 NBA 季后赛的胜负[10]。曾磐和朱安民通过建立多个指标来评估球队实力然后将 SVM 方法运用到季后赛胜负的预测当中，提出了球队的综合实力由球队的常规赛得分，球队的核心球员综合得分、主教练水平以及主客场因素四部分数据组成，取得了很好的预测效果[11]。半监督的学习算法在一些分类问题中有着很好的应用。半监督聚类作为半监督学习算法的一类，其相比于传统的聚类能够更好的利用已有的信息来达到聚类的效果，半监督聚类可以同时利用有标签和无标签的数据来进行聚类[12]。

本研究根据球队常规赛的战绩、球员的评价数据以及教练的历史执教数据构建出球队的综合评价数据。针对综合评价数据构建半监督聚类模型，季后赛第一轮对局的两只球队必须是一胜一负，这自然的就可以给出勿连约束。根据历史统计数据可以给出必连约束。构建模型以后可以将数据分为两类，根据历史数据可以给出每个类别的标签。预测结果表明季后赛第一轮的预测采用半监督聚类的方法具有很好的效果。

## 2. 数据组织

### 2.1. 数据来源

本文所用数据为 2020-21 赛季季后赛数据，来自于 NBA 权威的统计网站，包括 NBA 官方网站 https://www.sportingnews.com/us、娱乐与体育节目电视网 https://www.espn.com/nba/和体育统计网站 https://www.basketball-reference.com/。NBA 官方网站提供了球队常规赛期间每场比赛详细的统计数据，其中包括常规赛比赛得分、篮板、助攻等等一系列数据。ESPN 提供了球员的评价得分，其所用的 RPM (Real Plus-Minus)指标在球员评价方面认同度比较高，RPM 是考虑队友、对手以及其他因素基础上以每百回合中进攻和防守的净得分衡量球员在场上对球队表现影响的得分值，ESPN 提供的球员指标包括球员的进攻评价得分和防守评价得分，以及球员的综合评价得分等等。Basketball-reference 网站提供了教练当前赛季及历史赛季的执教表现，包括教练执教的年份、常规赛执教场次、常规赛执教胜利场次，常规赛执教失败场次，常规赛执教胜负场次之比、季后赛执教场次、季后赛执教胜利场次、季后赛执教失败场次，季后赛执教胜负比，季后赛执教获得分区冠军数以及季后赛执教获得总冠军数等。

### 2.2. 数据预处理

采集的原始数据为球队单场比赛的数据以及 NBA 历史上所有教练的执教数据和每个球员的评价值，不能综合的体现一只球队的实力。对于季后赛第一阶段的建模来说，可将原始数据整理汇总后分为三个大的方面来进行呈现：第一部分为球队的综合实力，主要运用球队常规赛期间的数据进行整理得到；第二部分为教练的执教水平，通过球队教练历史的执教数据整理获得；第三部分为球员实力的衡量，通过 ESPN 获取的球员赛季能力评价整理获得。以上三个方面能够较为全面反映出球队的实力水平。各个部分具体的内容见表 1。

本文所用数据中球队的综合实力是由 2020-21 赛季常规赛期间球队各场比赛数据汇总得到。球队的常规赛胜率是球队的胜场数与比赛总场次之比，主客场胜率 wh 和 la 分别是球队主客场胜利的场次与主客场场次的比值，以上三个指标在一定程度上反映了球队的综合实力。t2 到 t19 是球队赛场上数据的汇总平均值。场均得分 t2 是球队常规赛期间球队得分均值，反映了球队的得分能力，场均投篮相关数据命中数 t3、投篮数 t4 及命中率 t5 是球队的进攻能力。场均三分的相关数据是球队三分出手数 t6、命中数 t7 和命中率 t8 衡量球队三分线外的进攻能力，在现在 NBA 的“小球时代”，球队的三分投射能力是十分

**Table 1.** Team evaluation index
**表 1.** 球队评价指标

| 指标类别 | 指标名 | 指标含义 |
|---|---|---|
| 球队 | t1 | 球队常规赛胜率 |
| 球队 | t2 | 球队常规赛场均得分 |
| 球队 | t3 | 球队常规赛场均投篮命中数 |
| 球队 | t4 | 球队常规赛场均投篮数 |
| 球队 | t5 | 球队常规赛场均投篮命中率 |
| 球队 | t6 | 球队常规赛场均三分命中数 |
| 球队 | t7 | 球队常规赛场均三分出手数 |
| 球队 | t8 | 球队常规赛场均三分命中率 |
| 球队 | t9 | 球队常规赛场均罚球命中数 |
| 球队 | t10 | 球队常规赛场均罚球出手数 |
| 球队 | t11 | 球队常规赛场均罚球命中率 |
| 球队 | t12 | 球队常规赛场均前场篮板数 |
| 球队 | t13 | 球队常规赛场均后场篮板数 |
| 球队 | t14 | 球队常规赛场均篮板数 |
| 球队 | t15 | 球队常规赛场均助攻 |
| 球队 | t16 | 球队常规赛场均抢断 |
| 球队 | t17 | 球队常规赛场均盖帽 |
| 球队 | t18 | 球队常规赛场均失误 |
| 球队 | t19 | 球队常规赛场均犯规 |
| 球队 | Hca | 主场优势 |
| 球队 | wh | 球队常规赛主场胜率 |
| 球队 | la | 球队常规赛客场胜率 |
| 教练 | CRG | 教练常规赛执教场次 |
| 教练 | CRW/L | 教练常规赛执教胜负比 |
| 教练 | CPG | 教练季后赛执教场次 |
| 教练 | CPW/L | 教练季后赛执教胜负比 |
| 教练 | CPC | 教练获得总冠军次数 |
| 球员 | ORPM_num | 球员常规赛进攻表现占前百分之二十的人数 |
| 球员 | ORPM_total | 球员常规赛进攻表现前十的综合评分 |
| 球员 | DRPM_num | 球员常规赛防守表现占前百分之二十的人数 |
| 球员 | DRPM_total | 球员常规赛防守表现前十的综合评分 |
| 球员 | RPM_total | 球员常规赛综合表现前十的综合评分 |
| 球员 | WINS_total | 球队常规赛真实的正负值前十球员的综合评分 |
| 教练 | CRG | 教练常规赛执教场次 |
| 教练 | CRW/L | 教练常规赛执教胜负比 |

**Continued**

| | | |
|---|---|---|
| 教练 | CPG | 教练季后赛执教场次 |
| 教练 | CPW/L | 教练季后赛执教胜负比 |
| 教练 | CPC | 教练获得总冠军次数 |
| 球员 | ORPM_num | 球员常规赛进攻表现占前百分之二十的人数 |
| 球员 | DRPM_num | 球员常规赛防守表现占前百分之二十的人数 |
| 球员 | ORPM_total | 球员常规赛进攻表现前十的综合评分 |
| 球员 | RPM_total | 球员常规赛防守表现前十的综合评分 |
| 球员 | WINS_total | 球队常规赛真实的正负值前十球员的综合评分 |

重要的进攻得分手段。场均篮板 t12 是衡量球队对于篮板控制能力的指标，场均前场篮板 t13 是指球队进攻方向得到的篮板，场均后场篮板 t14 是指球队防守方向得到的篮板，获得前场篮板可以拥有二次进攻的机会，获得后场篮板往往意味着成功的防守。场均罚球相关数据 t9~t11 可以衡量球队突破及进攻篮下的能力。场均助攻 t15 是衡量得分前球的轮转次数。场均抢断 t16 和盖帽 t17 都属于衡量球队防守能力的指标。场均失误 t18 是球队失误数的平均值，场均犯规指标 t19 是球队每场犯规数的平均值，犯规和失误都是球队想要尽力减少的，球队过多的犯规和失误会导致对手得到更多的球权甚至是直接获得得分。

教练是球队的指挥者，每个教练都有自己的带队方式，例如有些侧重防守，有些侧重进攻，有些重视三分等等。历史常规赛执教场次 CRG 衡量教练常规赛经验水平，教练常规赛正负比 CRW/L 是教练常规赛执教场次中胜场数与负场数之比，比值越大代表胜场数越大，教练的常规赛执教水平越高。教练季后赛执教场次 CPG 是教练在季后赛的执教场次，教练季后赛执教胜负比 CPW/L 是教练季后赛执教场次中胜场数与负场数之比。季后赛期间两只球队的竞赛实行的是七局四胜制，这更加考验教练的排兵布阵能力，季后赛的比赛强度比较大，球员更容易疲劳甚至受伤，教练的季后赛场次和季后赛的胜负比能够体现教练的季后赛执教水平。教练获得总冠军次数 CPC 是教练过往赛季作为主教练获得总冠军的次数，获得总冠军的教练往往拥有更多的季后赛经验。NBA 的常规赛和季后赛是两种不同的竞赛方式，常规赛期间由于要磨合不同的阵容或者是让新加入球队的球员能够熟悉整个球队运行方式等原因，常规赛期间的教练成绩并不更够完全衡量教练的执教水平，因此可以加入季后赛的相关数据衡量教练的真实水平。

球员是球队真正上场比赛的人，球员的能力主要分为进攻能力和防守能力两大部分。ESPN 网站提供了球员常规赛结束以后其在整个赛季期间的进攻能力、防守能力以及综合能力评价得分，综合能力是指球员在场期间的效率值。一般来说一只球队的大名单中只能拥有十五人，不包括双向合同可以签订的两名球员，季后赛开始后上报整个球队球员的上场名单，名单中只能有 12 人，也就是季后赛每场比赛最多只能有 12 人上场轮换，一场比赛教练所使用的真正上场的球员往往在十人左右，而且球队中真正的主力球员在两或三人以内，约占队伍的百分之二十。球员常规赛进攻表现占前百分之二十的人数 ORPM_num 是将所有球员的进攻评分进行排序后统计球队进入前百分之二十的人数。球员常规赛防守表现占前百分之二十的人数 DRPM_num 是将所有球员的防守表现评分进行排序后统计球队中进入前百分之二十的人数。球员常规赛进攻表现前十的综合评分 ORPM_total 是将球队中所有球员的进攻表现得分排序，然后将排名前十的球员得分进行加总。球员常规赛防守表现前十的综合得分 RPM_tota 是将球员防守得分排序后取前十加总。球员常规赛综合表现评分 WINS_total 是将综合表现得分排序取前十加总。球员常规赛真实的正负值是指球员在场时净胜的效率值，将其排序后将排名前十的球员真实正负值进行加总得到球员常规赛真实的正负值前十的球员的综合评分。

## 3. 数据描述性统计

WorL 变量代表球队在 2020-21 赛季 NBA 季后赛第一轮比赛中的胜负。将季后赛第一轮建模的数据加入 WorL 变量后计算各个变量以及胜负变量的相关关系，统计结果见表 2。

**Table 2.** The correlation between first-round playoff data and team wins and losses
**表 2.** 季后赛第一轮数据与球队胜负相关性

| Person 相关系数 | 相关性 | 变量名 |
|---|---|---|
| $0.8 \le \lvert\rho\rvert < 1$ | 高度相关 | t1 |
| $0.5 \le \lvert\rho\rvert < 0.8$ | 中度相关 | t2, t5, t8, Hca, wh, la, RPM_total, WINS_total |
| $0.3 \le \lvert\rho\rvert < 0.5$ | 低度相关 | t3, t6, t11, t13, t14, t15, t19, ORPM_total, DRPM_num, DRPM_total |
| $\lvert\rho\rvert < 0.3$ | 基本不相关 | t4, t7, t9, t10, t12, t16, t17, t18, CRG, CRW/L, CPG, CPW/L, CPC, ORPM_num |

季后赛第一轮的胜负与球队常规赛胜率 t1 的相关系数为 0.81，球队常规赛的胜率可以体现球队的综合实力。在第一阶段的比赛中，由于采取的是东西部前八名进行对战，第一名对战第八名、第二名对战第七名、第三名对战第六名、第四名对战第五名，且排名前四的球队有着主场优势，所以球队常规赛战绩好的球队往往更能够获得季后赛第一轮的胜利。球队的主场优势 Hca 与季后赛第一轮胜负的相关关系为 0.63，球队常规赛主场胜率 wh 与季后赛第一轮的胜负的相关关系为 0.67，球队常规赛客场胜率 la 与胜负的相关关系为 0.55，球队的主客场会影响球队的胜负，球员在主场由于环境等影响往往会有更好的发挥，主场也会有裁判在裁决尺度上的优势，裁判的判决有很大的人为因素，主场作战的球队由于各种因素往往具有主场优势。

季后赛第一轮的胜负与球队常规赛场均得分 t2 的相关系数为 0.57、与球队常规赛场均投篮命中率 t5 的相关系数为 0.71、与球队场均三分命中率 t8 的相关系数为 0.58，球队常规赛的场均得分反映了球队的进攻能力，篮球是一个攻防相对的比赛，进攻能力和防守能力对于比赛的影响都很重要，当前时代的篮球和以前的篮球也有着很大的不同，自从勇士队崛起以后，小球时代成为了主流，三分投射能力对于球队的胜负非常关键，三分的投射能力能够拉开空间，这样球员在场上就会有更多的突破还有施展战术的机会，同时三分的投射会造成更多的长篮板，这样就能获得更多的进攻篮板进而获得二次进攻的机会。

球员常规赛综合表现前十的综合得分 DRPM_num 与胜负的相关系数为 0.72。球员综合表现前十的综合得分 DRPM_total 是对球员进行综合评价后将球队中综合排名前十的球员的综合得分进行加总的总分值，虽然简单的加总不能准确的反映球员在场上的表现，但是能够反映一只球队球员相对其他球队球员的实力，球员是球场上的关键人员，也是执行教练战术的人，球员的综合实力与球队能否取胜有着很大的关联，球员的正常发挥和正确执行战术是球队取胜的关键因素。季后赛第一轮的胜负与球队真实的正负值前十球员的加和的相关关系为 0.71，球队真实的正负值是指一个球员在球场上抛去其他因素以后在场的净胜分，也是反映球员实力的指标，对球队的胜负有着很大的影响。

将表 2 中季后赛第一轮数据与球队胜负相关系数高度相关和中度相关的变量绘制箱线图见图 1。从图 1 中可以看出季后赛第一轮胜利的球队其球队常规赛的胜率比较高，数据较为集中且为右偏分布。胜利的球队常规赛场均投篮命中数较多，总的投篮命中率以及三分命中率也较高。相比较输掉比赛的球队，胜利的一方无论是主场胜率还是客场胜率都比较高，其球队常规赛表现前十球员的综合评分与球队常规赛真实的正负值前十球员的综合评分也较高。

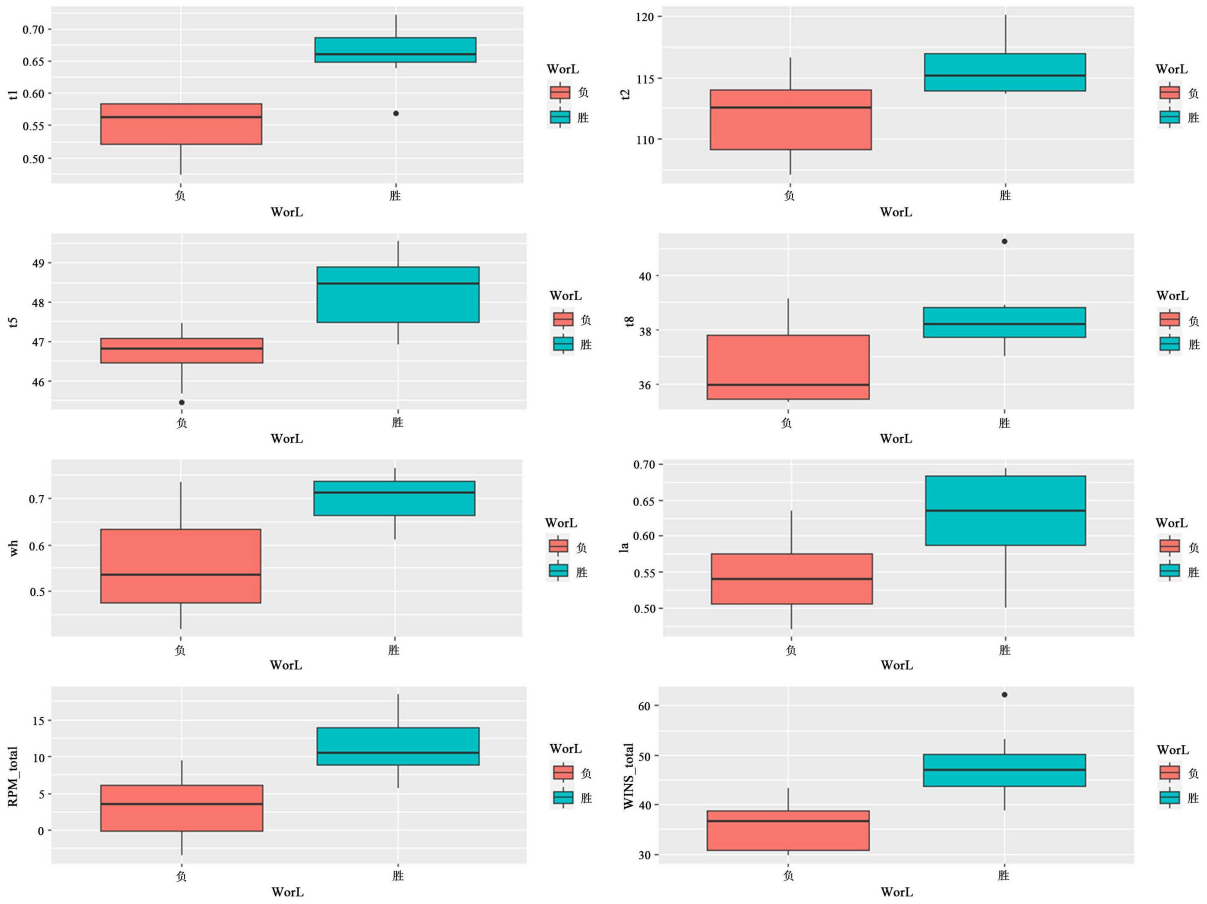


**Figure 1.** Boxplots of WorL and the variables of moderate correlation and high correlation
**图 1.** 季后赛第一轮胜负与中高度相关关系变量的箱线图

## 4. 半监督学习原理概述

半监督学习相对于传统的监督学习有着明显的优点和实用性。监督学习方法要求数据拥有全部的数据标签，这对于一些实际问题却是很难实现的，获取大量有标签的数据有时需要耗费大量的人力物力，也有时候获取标签是无法实现的，这时半监督学习的优点就体现出来了，半监督学习可以通过有标签的部分数据来提升学习性能，从而达到构建模型的目的。

半监督聚类是半监督学习的一种，通常分为两类。第一类是建立“必连”和“勿连”约束的半监督聚类，必连是指定某些样本属于一类，勿连是指定某些样本不属于一类。第二类是对于一组数据，只有其中一部分数据知道其标签，对于部分标签的数据训练建立模型，并且对数据中未有标签的那部分数据进行预测。典型的半监督聚类算法有 COP-Kmeans (Clustering Using Boosted Constrained k-Means)、改进的 LCOP-Kmeans (Linked Cop-Kmeans)、Seeded-Kmeans 和 MPCK-Means (Metric learning and pair-wise-constrained k-means)等，本文所使用的算法为 COP-Kmeans 和 MPCK-Means，并将其简写为 ckmeans 和 mpckm。算法流程如下：

算法 1：(基于约束的半监督聚类)

输入：样本集 $X=\{x_1,x_2,\cdots,x_N\}, x_i\in\mathbb{R}^d$；

必连约束集合$\mathcal{M}$；

勿连约束集合$\mathcal{C}$；

聚类簇数 $K$；

过程：

① 初始化种子 $S=U_{l=1}^{K}S_l$；

② 计算每一部分的均值 $\mu_h^{(0)}=\frac{1}{|S_h|}\sum_{x\in S_h}x,\ h=1,\cdots,K$；

③ 将每个数据点分配给 $h^*=\text{argmin}\left\|x-\mu_h^{(t)}\right\|^2$ 最小的簇；

④ 检测将 $x_i$ 划入聚类簇中是否会违背$\mathcal{M}$与$\mathcal{C}$中的约束；

⑤ 更新每个簇的均值，$\mu_h^{(t+1)}\leftarrow\frac{1}{X_h^{(t+1)}}\sum_{x\in X_h^{(t+1)}}x$；

⑥ 重复 3、4、5 步骤直至收敛；

⑦ 输出聚类结果。

## 5. 建立模型

聚类通常是一种无监督的算法，但是有时候我们能够知道一些先决的信息，这时可以利用这些先决信息建立半监督的学习算法从而获得更好的学习效果。在 NBA 季后赛第一轮的预测中一共有十六只球队共八组对局，我们所知道的先验信息就是两只球队之间必须是一胜一负，即作为对手的两只球队，分别属于不同的簇。

**Table 3.** The correlation between first-round playoff data and team wins and losses
**表 3.** 季后赛第一轮数据与球队胜负相关性

| 赛季 | 东部第一球队 | 是否获胜 | 西部第一球队 | 是否获胜 |
|---|---|---|---|---|
| 2011-12 | 芝加哥公牛 | 否 | 圣安东尼奥马刺 | 是 |
| 2012-13 | 迈阿密热火 | 是 | 俄克拉荷马城雷霆 | 是 |
| 2013-14 | 印第安纳步行者 | 是 | 圣安东尼奥马刺 | 是 |
| 2014-15 | 亚特兰大老鹰 | 是 | 金州勇士 | 是 |
| 2015-16 | 克利夫兰骑士 | 是 | 金州勇士 | 是 |
| 2016-17 | 波士顿凯尔特人 | 是 | 金州勇士 | 是 |
| 2017-18 | 多伦多猛龙 | 是 | 休斯顿火箭 | 是 |
| 2018-19 | 密尔沃基雄鹿 | 是 | 金州勇士 | 是 |
| 2019-20 | 密尔沃基雄鹿 | 是 | 洛杉矶湖人 | 是 |

通过表 3 中的数据可知，在 NBA2011 到 2019 共九个赛季季后赛第一轮东西部第一分别与各自分部的第八进行的 18 场比赛中，只有 2011-12 赛季芝加哥公牛在与第八名对局的过程中被逆转，也就是说在过去的九个赛季中只有 5.56%的概率会发生黑八，黑八指在 NBA 季后赛的第一轮系列赛中排名第一的球队在于排名第八的球队的系列赛中输掉比赛没有晋级，在 NBA 的近些年的历史上也很少有黑八现象的发生，季后赛第一轮的比赛情况是由球队的常规赛的战绩决定的，在常规赛的比赛过程中，球队的实力虽然不能完全体现在球队的统计数据中，但是排名靠前的球队在实力上还是明显强于排名靠后的球队，东西部第一与第八的实力差距明显。

由于 2020-21 赛季首次采用附加赛的赛制，也就是进入东部与西部中的第 7、8 名是由各自分区的第

7、8、9、10 竞争得到的，具体的赛制为常规赛战绩排名的第 7 名与第 8 名角逐，胜者获得第 7 名的位次，同时进行的还有第 9 与第 10 名的角逐，第 7 名与第 8 名角逐中的败者与第 9 与第 10 名的角逐中的胜者角逐第 8 名。以上的附加赛赛制对于 2020-21 赛季以前的比赛没有影响，对于当前赛季的影响是西部联盟的金州勇士被淘汰，孟菲斯灰熊获得了西部第八进入了总决赛。附加赛结束后的成绩排名见表 4，表 4 中括号里面为球队的英文名称缩写。

**Table 4.** Regular season ranking
**表 4.** 常规赛排名

| 排名 | 东部联盟 | 西部联盟 |
|---|---|---|
| 1 | 费城 76 人(PHI) | 犹他爵士(UTA) |
| 2 | 布鲁克林篮网(BKN) | 菲尼克斯太阳(PHX) |
| 3 | 密尔沃基雄鹿(MIL) | 丹佛掘金(DEN) |
| 4 | 纽约尼克斯(NYK) | 洛杉矶快船(LAC) |
| 5 | 亚特兰大老鹰(ATL) | 达拉斯独行侠(DAL) |
| 6 | 迈阿密热火(MIA) | 波特兰开拓者(POR) |
| 7 | 波士顿凯尔特人(BOS) | 洛杉矶湖人(LAL) |
| 8 | 华盛顿奇才(WAS) | 孟菲斯灰熊(MEM) |

在半监督聚类模型的构建构成中，根据表 3 中的历史数据，可以给出“必连”(must-link)：东部第一和西部第一在一个簇中，也即费城 76 人和犹他爵士在一个簇中。模型中的“勿连”(cannot-link)是指每两只对局的球队在不同的簇中，季后赛第一轮采取的对局方式为第一名对阵第八名、第二名对阵第七名、第三名对阵第六名、第四名对阵第五名，东西部两个联盟分别进行比赛，这也构成了勿连约束。勿连约束$\mathcal{M}$与勿连约束$\mathcal{C}$如下，设定聚类簇数 $K = 2$ 构建模型。

必连约束集合$\mathcal{M}$ = {(费城 76 人，犹他爵士)}。

勿连约束集合$\mathcal{C}$ = {(费城 76 人，华盛顿奇才)，(布鲁克林篮网，波士顿凯尔特人)，(密尔沃基雄鹿，迈阿密热火)，(纽约尼克斯，亚特兰大老鹰)，(犹他爵士，孟菲斯灰熊)，(菲尼克斯太阳，洛杉矶湖人)，(丹佛掘金，达拉斯独行侠)，(洛杉矶快船，达拉斯独行侠)}。

R 语言的 SSLR 包提供了半监督聚类算法。通过 SSLR 包构建 ckmeams 和 mpckm，然后对于分类的结果定义东部第一和西部第一所在的类为胜者组，另一组相应的为败者组，预测结果见表 5，表 5 中显示的是表 4 中球队的英文名缩写。ckmeans 计算预测准确率为 16/16，mpckm 预测的准确率为 14/16，每两只球队为一场对局，将预测结果转换为对局形式再次计算准确率，ckmeans 计算预测准确率为 8/8，mpckm 预测的准确率为 7/8。

半监督的学习算法在 NBA 季后赛第一轮的预测中有很好的适用性，第一轮的比赛对于半监督聚类的模型可以给出勿连和必连的限制条件，同时能够给予划分出来的类一个标签。这些前提条件能够完美满足半监督聚类的设定。第一轮比赛中对局的双方很多比赛都有着很明显的差距，第一轮的对局中是东西部第一与第八进行比赛，第二与第七，第三与第六，第四与第五，除东西部第四与第五的比赛外，其余实力差距过大，这样在做聚类进行二分类时会有更好的结果，在这其中对于球队的评价也是一个重要的方面。

**Table 5.** Semi supervised prediction results
**表 5.** 半监督预测结果

| 预测方法 | PHI | WAS | NYK | ATL | MIL | MIA | BKN | BOS | UTA | MEM | LAC | DAL | DEN | POR | PHX | LAL | 预测准确率 | 对局准确率 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ckmeans | 胜 | 负 | 负 | 胜 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 16/16 | 8/8 |
| mpckm | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 胜 | 负 | 14/16 | 7/8 |

## 6. 结论与展望

季后赛的预测实际是一个二分类问题，本文将半监督聚类算法应用在 NBA 季后赛第一轮的预测中，其中构造球队相对实力的准确评价数据是建立模型的基础，只有正确的或者是相对正确的对于球队给出球队的综合能力评价指标与其值才能在一个实用的模型中得到较好的结果。本文所使用的数据从球队、教练和球员三个大的方面进行呈现，球队部分的数据主要从球队当前赛季常规赛数据组织得到。教练部分的数据是截至到当前预测季后赛开始前球队教练的执教数据，球员部分的数据是球员当前赛季综合评分整理得到的，针对整理的数据建立半监督模型进行预测，预测结果表明研究思路有着很好的效果。

本研究没有将数据进行降维处理，同时许多的指标的建立也来自于历史经验。对于进一步的研究，针对本文的数据也可以尝试使用半监督聚类中使用部分标签数据对另外一些数据进行预测的算法，同时也可以探索将此数据运用在 NBA 后三轮的预测当中。

## 参考文献


[1] Chen, V.C., Kim, S.B., Oztekin, A. and Sundaramoorthi, D. (2018) Preface: Data Mining and Analytics. *Annals of Operations Research*, **263**, 1-3. https://doi.org/10.1007/s10479-018-2787-1

[2] Turban, E, Sharda, R. and Delen, D. (2010) Decision Support and Business Intelligence Systems. 9th Edition, Prentice Hall, New Jersey, 4-19.

[3] Senderovich, A., Shleyfman, A., Weidlich, M., Gal, A. and Mandelbaum, A. (2018) To Aggregate or to Eliminate? Optimal Model Simplification for Improved Process Performance Prediction. *Information Systems*, **78**, 96-111. https://doi.org/10.1016/j.is.2018.04.003

[4] Gerrard, B. (2016) Moneyball and the Role of Sports Analytics: A Decision-Theoretic Perspective. *North American Society for Sport Management Conference*, Orlando, 31 May-4 June 2016, 108-109.

[5] 王明新. 2015-2016 赛季 NBA 总决赛克利夫兰骑士队制胜因素分析[D]: [硕士学位论文]. 新乡: 河南师范大学, 2017: 49-50.

[6] 靳勇, 李永辉, 李丽. 2005-2006 赛季 NBA 总决赛制胜因素探析[J]. 哈尔滨体育学院学报, 2006(6): 118-119.

[7] 邱胜, 段重阳, 陈征. NBA 季后赛成绩分析及预测: Logistic 和 Bayes 模型[J]. 统计教育, 2010(10): 46-51.

[8] Sports Reference (2022) Glossary. https://www.basketball-reference.com/about/glossary.html

[9] Hsu, P.H., Galsanbadam, S., Yang, J.S. and Yang, C.Y. (2018) Evaluating Machine Learning Varieties for NBA Players' Winning Contribution. 2018 *International Conference on System Science and Engineering* (*ICSSE*), New Taipei, 28-30 June 2018, 1-6. https://doi.org/10.1109/ICSSE.2018.8520017

[10] Cheng, G., Zhang, Z., Kyebambe, M.N. and Kimbugwe, N. (2016) Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle. *Entropy*, **18**, Article 450. https://doi.org/10.3390/e18120450

[11] 曾磐, 朱安民. 基于支持向量机的 NBA 季后赛预测方法[J]. 深圳大学学报(理工版), 2016, 33(1): 62-71.

[12] 秦悦. 成对约束半监督聚类算法研究[D]: [硕士学位论文]. 徐州: 中国矿业大学, 2020.