**2022年卡塔尔世界杯各国球队分析及胜率预测**

符亦铭 2020012851 未央水木01

# 摘要

本研究首先在原数据集中选择反映球队表现的核心变量，对数据集进行预处理，得到所有球队整体表现数据。对于球队水平，采用因子分析，探究影响球队表现的潜在因素；采用典型相关性分析，探究球队不同阵线间的联系；采用聚类分析，探究表现相似的球队。对于比赛水平，采用回归分析，探究影响比赛结果的主要因素，并结合实际案例，对比赛胜率进行预测，最终得到较为合理的结果。

**目录**

[摘要 1](#_Toc138546533)

[1 研究背景及问题 2](#_Toc138546534)

[1.1 背景介绍 2](#_Toc138546535)

[1.2 研究问题 2](#_Toc138546536)

[2 数据集介绍 3](#_Toc138546537)

[2.1 总体介绍 3](#_Toc138546538)

[2.2 变量选择 3](#_Toc138546539)

[2.3 数据预处理 4](#_Toc138546540)

[3 球队水平分析 5](#_Toc138546541)

[3.1 因子分析FA：影响球队表现的潜在变量 5](#_Toc138546542)

[3.2 典型相关分析CCA：球队不同阵线间的联系 7](#_Toc138546543)

[3.3 聚类分析CA：表现较为相似的球队 8](#_Toc138546544)

[4 比赛水平分析 10](#_Toc138546545)

[4.1 回归Regression：影响比赛结果的主要因素 10](#_Toc138546546)

[4.2 回归Regression：比赛胜率预测及实例分析 11](#_Toc138546547)

[5 总结与展望 12](#_Toc138546548)

[5.1 研究结论 12](#_Toc138546549)

[5.2 限制及改进 12](#_Toc138546550)

# 1 研究背景及问题

## 背景介绍

在2022年冬季，来自不同地区、不同大洲的共32支球队，在卡塔尔开启了世界杯正赛的征程。最终阿根廷队在首战爆冷负于沙特阿拉伯队后，全队都激发出昂扬斗志，一路力克强敌过关斩将，在决赛中点球大战击败法国队，夺得2022年卡塔尔世界杯冠军。阿根廷队长梅西也一战封王，举起大力神杯的同时也梅开二度，再次获得世界杯金球奖，得到国际足联FIFA官方认证的球王封号。此外，本届世界杯也不乏精彩赛事，例如亚洲劲旅日本队爆冷连克德国队与西班牙队，上届亚军克罗地亚队点球大战接连取胜，欧洲黑马摩洛哥队一路逆袭闯进四强等。而这届精彩纷呈的大赛背后，又体现出怎样的数据规律，这有待我们深入探索。

## 研究问题

基于集合逻辑的视角，足球等体育赛事具有其独特之处：足球比赛可以看作一个二元联系集（此处的比赛指两队之间的比赛），分别对应两支球队（实体）和一场比赛（联系），因此对于足球比赛的分析可以从如下两个维度/水平进行：

**1.球队水平：**每支球队在比赛中都有对应的技术统计，将一支球队所进行过的每场比赛的技术统计进行综合，可以得到该球队在这届赛事中的**总体表现**。对每支球队都进行如上所述的操作，即可得到所有参赛球队各自的表现。

（注：本研究中将以“表现”来指代球队核心数据统计的指标，包括原始的统计数据与经处理得到的数据；而“水平”则指代考虑的维度/level，使用方法类似于“总体水平”，而非球队的技术能力及表现，此处加以说明，避免引起混淆。）

**2.比赛水平：**由于比赛本身是一个二元联系集，因此比赛双方各自的表现与球队特点都将会影响比赛结果。为建立实际的预测模型，并且结合足球比赛的现实意义（足球比赛中强调发挥己方优势与针对敌方劣势），本研究中以比赛双方球队的**表现差异**来作为衡量比赛结果的因素，也即是回归模型中的自变量。

针对上述两个研究水平，本研究提出了5个主要的研究问题，并将通过如下对应的方法进行研究：

**球队水平：**

1.因子分析FA：是否存在影响球队表现的潜在变量？

2.典型相关分析CCA：球队不同阵线间存在什么样的联系？

3.聚类分析CA：那些球队的表现较为相似

**比赛水平：**

4.回归Regression：影响比赛结果的主要因素有哪些？

5.回归Regression：如何进行比赛胜率预测（并结合实例进行分析）

# 2 数据集介绍

## 2.1 总体介绍

本研究的数据集来自网络学堂的备选数据集二，数据集中并没有缺失值，也不存在重复值/异常值，无需再进行合理化处理。数据集中的每个样本为一场比赛的信息，记录了比赛的对阵双方、时间、比赛性质（小组赛/淘汰赛）、比赛结果以及两支球队的技术统计，共有64个样本（对应48场小组赛与16场淘汰赛）和88个变量。值得注意的是，csv文件最后两行是对于变量的描述，在代码中需要注意取前64行，不要将最后两行读入。

## 2.2 变量选择

考虑到此处变量个数88大于样本数量64，且变量中存在大量平行变量（左/中/右路数据“left/central/right channel team1/2”等）、无用变量（比赛日期“date”等），且许多次要变量（较为细化的数据统计，例如尝试打破阵线次数“attempted line breaks team1/2”等）只是描述球队在某一具体细节方面的表现，难以直观体现球队的表现，因此需要对变量进行选择。

结合足球比赛常用的技术统计，从数据集中提取如下13个变量作为分析考虑的核心变量：控球率、进球数、射门数、射正率（射正数/射门数）、失球数、助攻数、犯规数、传球数、传球成功率（传球成功数/传球数）、扑救率（扑救数/对方射门数）、造成失误数、压迫数、净胜球（进球数-失球数）。其中由于净胜球数能够直接反应比赛结果（净胜球＞0表示获胜，净胜球=0表示平局，净胜球＜0表示落败），因此将净胜球作为结果变量考虑，与其他变量进行区分。

注1：此处的扑救数包括防守方在进攻方射门前进行的破坏，导致进攻方并未形成射门，因此扑救数通常会大于对方射门数，也即扑救率通常会大于1。此处将扑救率理解为球队对于对方射门的化解能力，其值越大表示化解能力越强。

注2：原数据中的“fouls against team1/2”变量表示球队受到的犯规，转化为球队犯规数时需要使用对方对应的变量。

研究中选择的变量如下表1所示，其中还加上了球队名称。

表 1 变量选择表

|  |  |  |
| --- | --- | --- |
| **变量名称** | **原变量名** | **取值类型** |
| 球队 | team | 字符型 |
| 控球率 | possession | 数值型（0～1） |
| 进球数 | number of goals | 整数型 |
| 射门数 | total attempts | 整数型 |
| 射正率 | on target attempts/total attempts | 数值型（0～1） |
| 失球数 | conceded | 整数型 |
| 助攻数 | assists | 整数型 |
| 犯规数 | fouls (against) | 整数型 |
| 传球数 | passes | 整数型 |
| 传球成功率 | passes completed/passes | 数值型（0～1） |
| 扑救率 | goal preventions/total attempts | 数值型 |
| 造成失误数 | forced turnovers | 整数型 |
| 压迫数 | defensive pressures applied | 整数型 |
| 净胜球 | number of goals-conceded | 整数型 |

## 2.3 数据预处理

根据上述变量选择，从原数据集中提取核心变量（需要包括team 1和team 2的对应变量），得到简化后的数据集。此处的每个样本仍然是一场比赛的信息，为获取每支球队的总体表现水平，需要对每场比赛信息进行处理。对于每场比赛，需要分别提取参赛两球队的核心数据，记录在数据框（python代码中为data\_df）中。随后对于数据框按照球队名称进行分组（python代码中为groupby），对结果求平均值（由于每支球队的比赛场数不尽相同，因此取平均值以便球队间进行比较）,即可得到每支球队在本届世界杯中各项核心数据的平均值，代码结果展示的前10支球队（按照球队名称排序）数据如下图1所示。



图 1 前10支球队数据展示图

对球队数据进行描述（python代码中为describe），结果如下图2所示。



图 2 球队数据描述统计结果图

# 3 球队水平分析

## 3.1 因子分析FA：影响球队表现的潜在变量

为探究有哪些潜在变量会影响球队表现，此处进行因子分析。

**因子个数确定**

利用python中的factor\_analyzer库，可得到如下图3所示的崖底碎石图。经计算可得（python代码中的get\_factor\_variance，返回结果的第三个数组即为各因子累计解释方差），三个因子可以解释64.16%的方差（这个结果可以接受）。考虑到因子的解释性，如果因子个数太多，则难以对球队进行解释，如果因子个数太少，则只能解释至多54.57%的方差（两个因子的情况），因此选用三个因子进行因子分析，后续会说明三个因子具有较好的解释性。

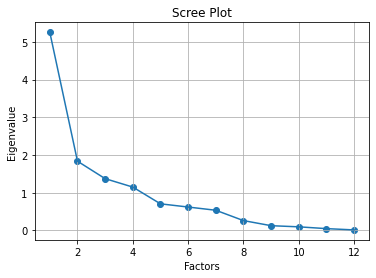


图 3 因子分析的崖底碎石图

**因子载荷可视化**

为获得更好的可解释性，对因子载荷矩阵进行旋转，此处主要考虑两种旋转方式：正交方差极大旋转（varimax）与斜交方差极大旋转（oblimax），使用seaborn库中的heatmap函数，可以将因子载荷可视化，如下图4与图5所示。

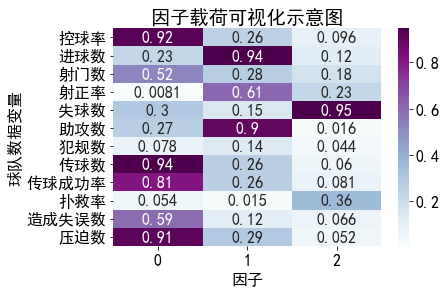


图 4 varimax旋转法的因子载荷可视化示意图

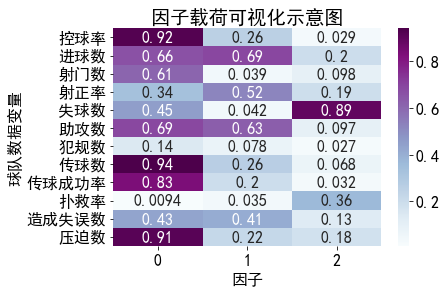


图 5 oblimax旋转法的因子载荷可视化示意图

通过对比上图4与上图5可以发现，相较于oblimax方法的结果而言，varimax方法的结果更好，体现在第一个因子（因子0）的组内差异较大，而第二、三个因子（因子1、2）的共性较大，因此选择varimax方法得到的结果。

**因子结果解释**

结合实际足球比赛的知识，从上图4中可以发现，因子0中主要的变量为控球率、传球数、传球成功率、造成失误数、压迫数，可以解释为球队的**中场能力**（组织、拦截能力）；因子1中主要的变量为进球数、射正率、助攻数，可以解释为球队的**前场能力**（进攻能力）；因子2中主要的变量为失球数、扑救率，可以解释为球队的**后场能力**（防守能力）。

注：因子2中扑救率的载荷实际为负数（绘图结果取绝对值进行展示，其余参与因子解释的载荷均为正数），因此失球率与扑救率反相关，故可以解释为球队的防守能力。而由于失球率越高防守能力应该越差，因此这里的防守能力是相反的，实际衡量防守能力（因子2得分）时，应当取相反数进行考虑与分析。

## 3.2 典型相关分析CCA：球队不同阵线间的联系

根据因子分析结果，可以将球队的数据变量大致分出三类，分别对应球队的前场、中场、后场能力。为考虑球队不同阵线间的联系，此处进行典型相关分析。考虑到实际足球比赛中，前场与后场的直接联系较少，因此主要考虑前场与中场、中场与后场的联系。

**前场与中场联系**

对所有球队的前场与中场数据进行CCA分析，得到Helio图如下图6所示。

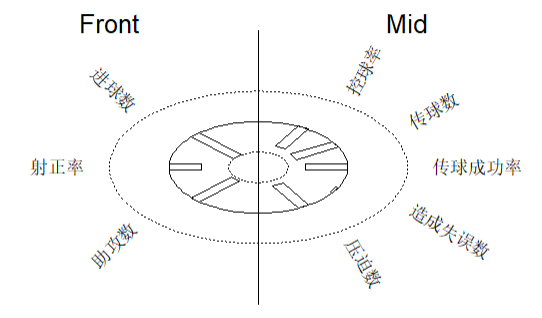


图 6 前场与中场CCA结果示意图

从上图6中可以发现，前场数据可解释为进球能力（进球数、助攻数），中场数据可解释为综合支配球能力（控球率、传球数、传球成功率、压迫数），说明前场的**进球能力**和中场的**综合支配球能力**有很强相关性，相关系数约为0.5856。

**中场与后场联系**

对所有球队的中场与后场数据进行CCA分析，得到Helio图如下图7所示。

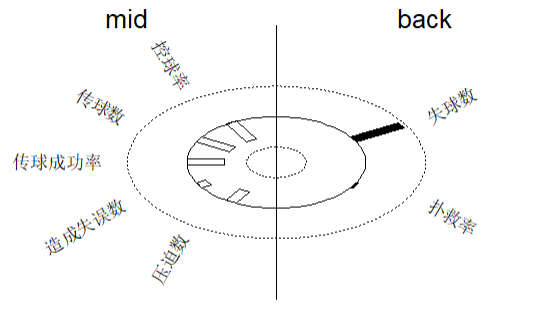


图 7 中场与后场CCA结果示意图

从上图7中可以发现，中场数据可解释为控球能力（控球率、传球数、传球成功率），后场数据可解释为防失球能力（失球数，此处为反相关），说明中场的**控球能力**和后场的**防失球能力**有很强相关性，相关系数约为0.6724。

## 3.3 聚类分析CA：表现较为相似的球队

为探究球队表现较为相似的球队，此处进行聚类分析，使用因子分析后的得分作为球队表现数据，分别对应球队的中场、前场、后场（取相反数）能力。

**聚类数目**

从下图8中可以看出，当聚类数目k=4时，总残差平方和已经较小。

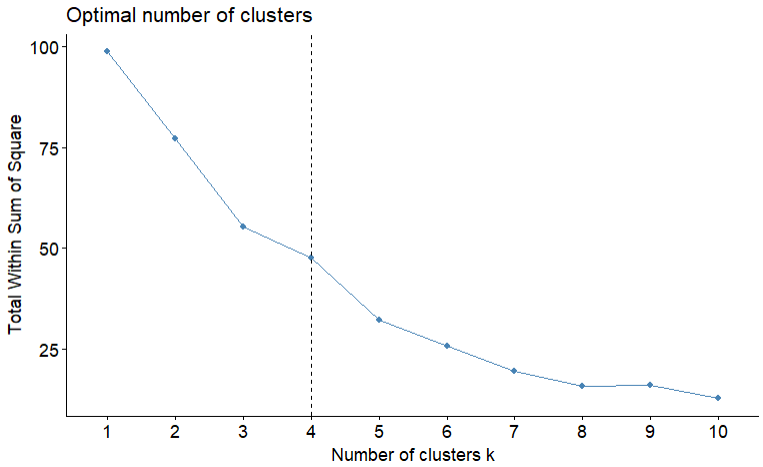


图 8 总残差平方和与聚类数目关系

**先聚类再降维**

先对三维数据进行Kmeans聚类，再用PCA降维成二维数据，如下图9所示。

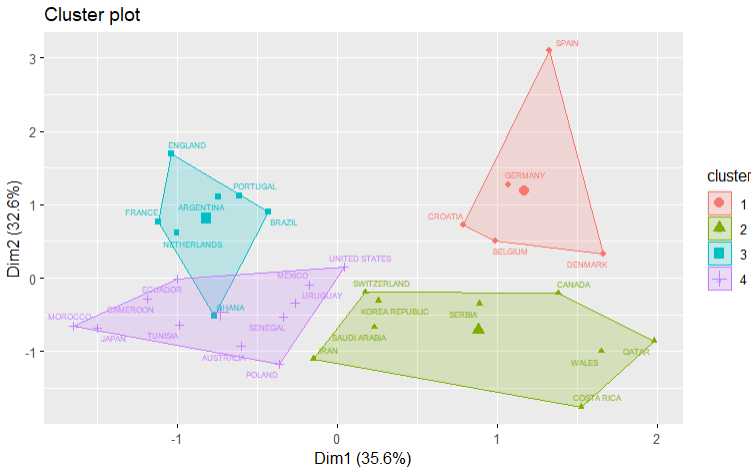


图 9 先聚类再降维结果示意图

从上图9中可发现，左上角的国家（如阿根廷、法国等）为实力较强且表现好的球队，右上角的国家（如德国、西班牙等）为实力较强但表现较差的球队，左下角的国家（如日本、摩洛哥等）为实力较弱但表现较好的球队，右下角的国家（如加拿大、卡塔尔等）为实力较弱且表现较差的球队。聚类结果整体合理且有较好解释性，但左上角类中混有加纳（不准确项），因此考虑不同的聚类过程。

**先降维再聚类**

先对三维数据用PCA降维成二维数据，再进行Kmeans聚类，如下图10所示。

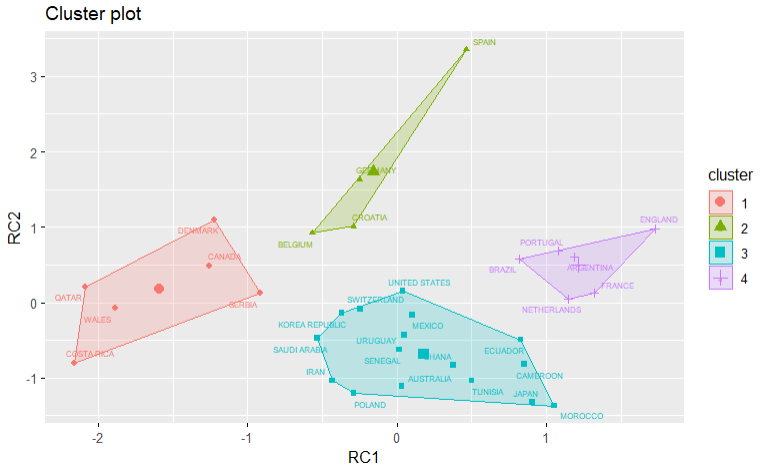


图 10 先降维再聚类结果示意图

上图10的结果仍可用相同的方式对四类进行解释，同时不准确项加纳的问题也得以解决，整体结果与之前结果类似，具有良好的可解释性。（注：由于大部分弱队的表现优劣难以衡量，此处最主要考虑强队的表现优劣与分类情况。）

# 4 比赛水平分析

## 4.1 回归Regression：影响比赛结果的主要因素

为探究影响比赛结果的主要因素并对比赛胜率进行预测，此处进行回归分析。比赛中净胜球可完全确定比赛结果，其大小可以相对反映两队实力差距，从而近似给出比赛胜率，此处对净胜球数进行Sigmoid变换，得到相对胜率（可以解读为获胜的概率）：。（系数防止胜率随净胜球变化太快。）

考虑到比赛为两球队之间的一种对比，因此使用两球队在前/中/后场的能力差值（分别对应、、）作为决定获胜概率的因素，对因子得分数据做差并标准化，得到对应数据表格，进行线性拟合得到的公式为：，回归诊断图如下图11所示。

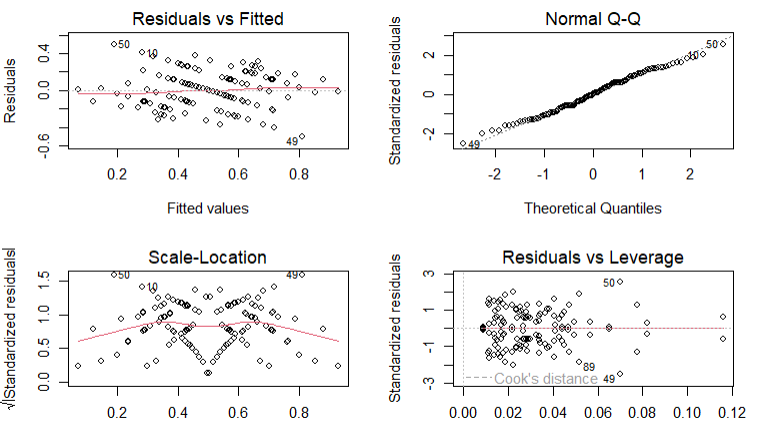


图 11 线性回归诊断图

从上图11中可发现，残差拟合图近似直线，Q-Q图近似线性，位置尺度图两侧点不太符合随机分布，残差杠杆图无明显离群点，因此可以判断拟合满足线性、正态性假设（较为合理），但不太满足同方差性。从回归表达式可以发现，**影响比赛结果**的最主要因素为前场差距，随后为后场差距，最后为中场差距。（注：此处使用全部比赛数据，也可使用小组赛/淘汰赛数据分别拟合，结果相似。）

## 4.2 回归Regression：比赛胜率预测及实例分析

使用之前的到的回归表达式，可以对比赛胜率进行预测。此处考虑阿根廷队经历的四场比赛（1:2爆冷沙特、2:2点球大战力克荷兰、3:0轻取克罗地亚、3:3点球险胜法国），预测得到的比赛胜率如下图12所示。



图 12阿根廷队四场比赛胜率预测

从上图12中可发现，阿根廷队对沙特的胜率很高，对荷兰略有优势，对克罗地亚优势较大，对法国稍显劣势，与实际比赛情况相符合，说明回归模型较为合理（由篇幅限制，其它比赛胜率预测的实例分析此处省略，几乎全部符合实际。）

# 5 总结与展望

## 5.1 研究结论

**问题1：**存在影响球队表现的潜在变量，可以理解为球队前/中/后场能力。

**问题2：**球队前场的进球能力与中场的综合支配球能力、中场的综合支配球能力与后场的防失球能力之间存在较强的相关性。

**问题3：**球队按照实力与表现大致分为四类，可解释性较好，尤其对于强队。

**问题4：**对比赛胜率的影响程度依次为：前场差距＞后场差距＞中场差距。

**问题5：**基于回归模型可以预测比赛胜率，与实际案例具有极强的一致性。

## 5.2 限制及改进

**限制：**足球比赛的数据集中样本较少而变量很多，因此对于偶然性可能较为敏感（相对权重较大），同时许多变量并未实际参与数据分析的过程。

**改进：**可以结合往届世界杯数据，对参赛国家进行纵向分析，相当于增加样本数量；可以考虑如何处理未使用的变量，将其组合为新的变量反映球队表现，从而更好地利用所有数据，呈现出更为全面的球队表现及更为准确的比赛胜率预测。