**体育大数据背景下的NBA球员分析**

**作者**：罗秋红，陈希，易沛，程合晴，冯永

**单位：**重庆大学，计算机学院

**案例版权：**该案例归重庆大学计算机学院所有

**涉及的知识点：**体育大数据，数据采集，数据预处理，聚类分析，可视化

**案例来源及案例真实性情况：**该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 伴随着大数据时代的到来，数据被视为了一种新的生产要素和创新驱动力。“大数据”概念自提出以来，已迅速成为传统行业创新发展的重要参考。正值全民体育时代的今天，将体育与大数据融合已是大势所趋。拥有强力数据支持的体育赛事与体育组织，相对比其他竞争者来说有难以替代的优势。以大数据背景下的NBA赛事为例，先进系统的大数据分析把单场比赛比分、各场馆上座人数、各位球员得分这些零散的点串连成线，变为场上成熟的战术模式和科学的商业发展模型，强有力的数据支持为联赛指明了更为清晰的发展方向与策略。本案例探讨了大数据背景下的NBA球员分析，主要分为三个部分：（1）通过网络爬虫进行数据采集，获取球员数据信息，并对数据进行缺失值填充和特征清洗；（2）对球员进行聚类分析，从而对球员的综合素质进行评判。（3）进行球员数据聚类分析结果的可视化。

**关键词：**体育大数据，数据采集，数据预处理，聚类分析，可视化

# 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为选取得分(PTS)、篮板(TRB)、助攻失误比(ATR)对球员综合素质进行刻画。基于以上，需要引导学生进行的主要内容有：（1）通过网络爬虫进行数据采集，获取球员数据信息；（2）对球员进行聚类分析，从而对球员的综合素质进行评判。（3）进行球员数据聚类分析结果的可视化。

# 2 背景介绍

伴随着大数据时代的到来，数据被视为了一种新的生产要素和创新驱动力。“大数据”概念自提出以来，已迅速成为传统行业创新发展的重要参考。正值全民体育时代的今天，将体育与大数据融合已是大势所趋。拥有强力数据支持的体育赛事与体育组织，相对比其他竞争者来说有难以替代的优势。以大数据背景下的NBA赛事为例，先进系统的大数据分析把单场比赛比分、各场馆上座人数、各位球员得分这些零散的点串连成线，变为场上成熟的战术模式和科学的商业发展模型，强有力的数据支持为联赛指明了更为清晰的发展方向与策略。在本案例中，我们选取得分(PTS)、篮板(TRB)、助攻失误比(ATR)使用这三个指标对球员综合素质进行评估。

**3 内容**

本文主要分为四个部分，分别为球员数据数据采集、数据预处理、球员聚类分析以及数据可视化。

3.1 数据采集

本次大数据项目的数据来源于网站stat-nba.com，网站里面的数据齐全，包含球员的单场比赛数据、生涯数据、赛季数据、常规赛数据以及季后赛数据等。为了从该网站获得本次项目的实验数据，我们采取了网络爬虫进行数据获取。爬取的流程如图1所示。首先进入网站查看网站的源码，获得数据所在代码位置，并进行解析；得到源代码的位置后采用python爬虫库requests和解析数据解析库BeautifulSoup；在爬取的过程中同时使用python的re，numpy，pandas等库进行数据的清洗和保存。

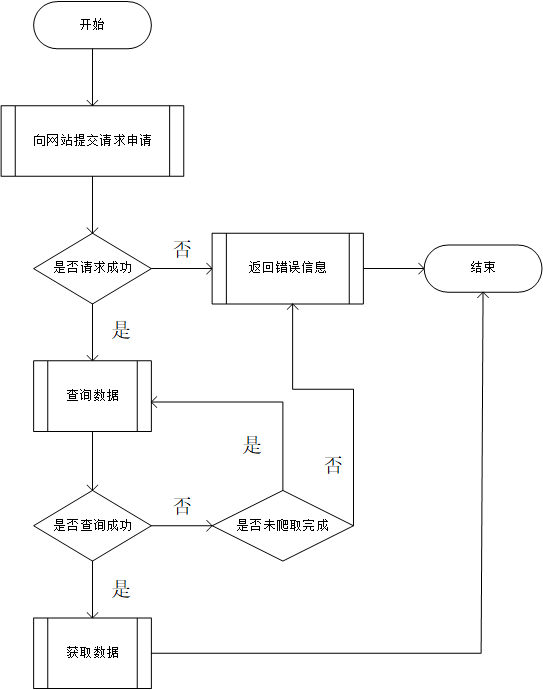
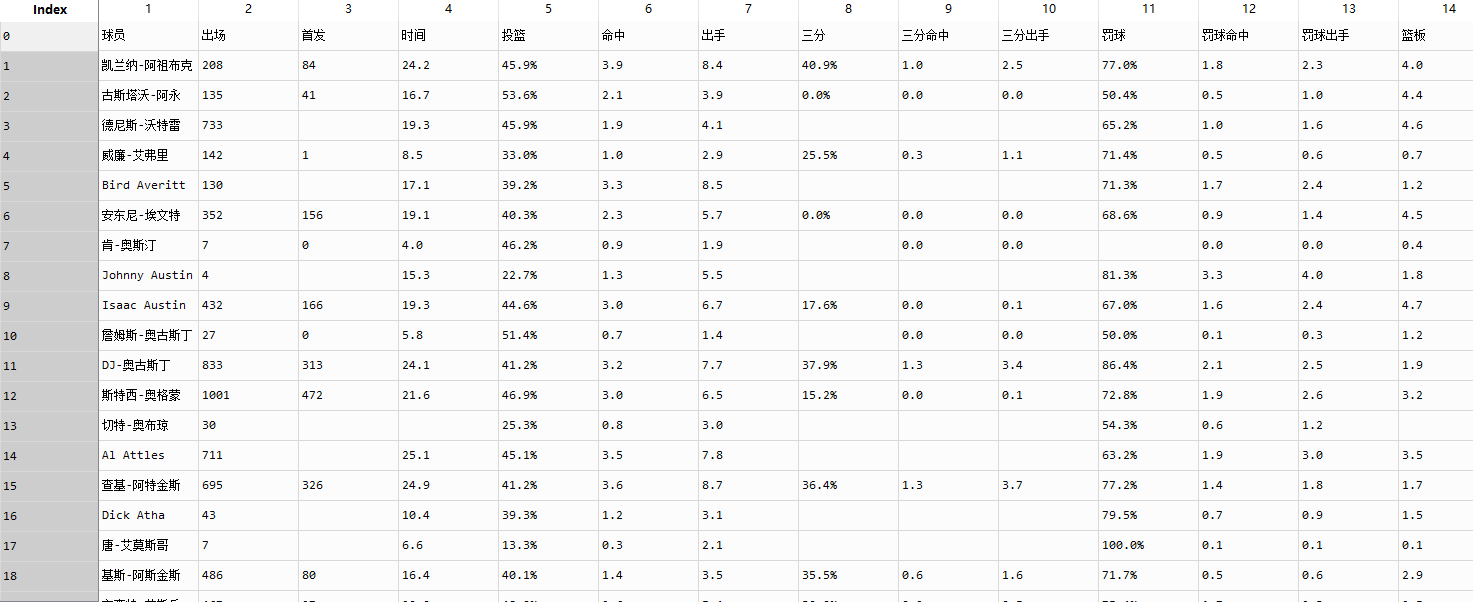


图1 爬虫流程

收集到的数据格式如图2所示，数据中主要是由：出场、首发、时间、投篮、命中、出手、三分、三分命中、三分出手、罚球、罚球命中、罚球出手、篮板、前场、后场、助攻、抢断、盖帽、失误、犯规、得分、胜、负等23个球员数据特征构成的数据表。

图2 数据格式

3.2 数据预处理

（1）特征删除

首先对数据特征进行观察，出场、首发、时间特征对本次球员的数据分析不会有太大的重要，可以删除；同时我们是对球员的数据进行分析，分析球员的投篮能力，助攻能力，篮板能力，所以胜，负对于球员分析没有太大的影响，胜负是对于球队的，个人数据很优秀不代表胜的场次就多，同样的球队的胜场很高，但假如是板凳球员，那胜负对于球员是都是常规球员的功劳，所以可以删除这两个特征。

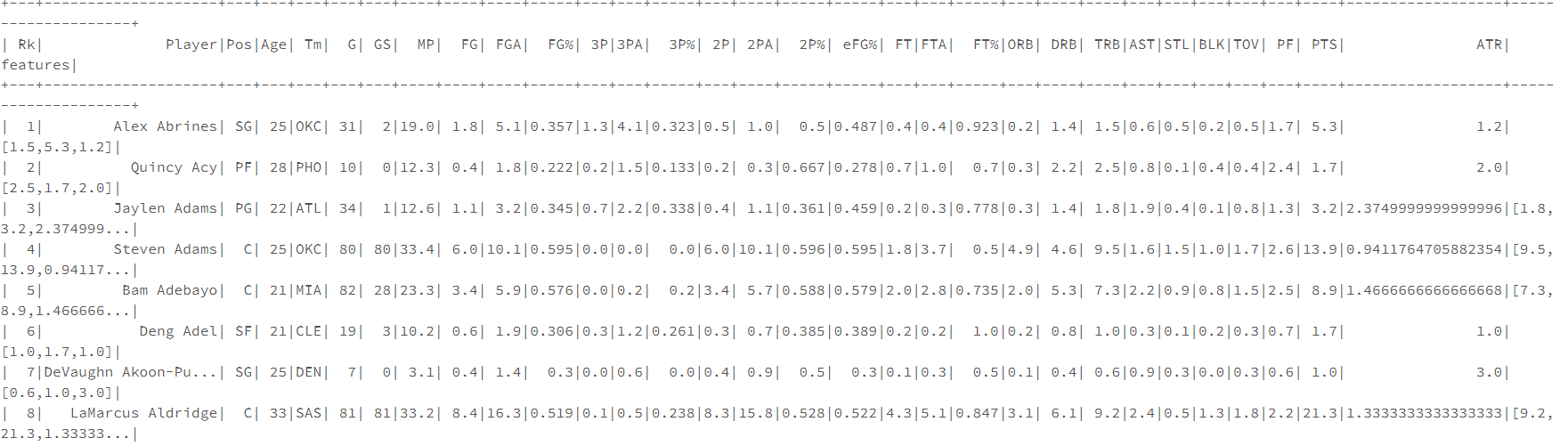
（2）缺失数据处理

通过对获取的数据进行观察可以发现存在很多缺失数据，而空值对最后的数据分析是会造成很大的影响。对于缺失的数据，每个人的数据之间是没有相互影响的，即每一列的特征之间的数据是没有相互关系，这对我们的缺失值填充带来了很大方便。以三分球数据特征缺失为例，我们可以用0进行填充。

（3）数据集上传

通过上述的简单分析我们可以对所爬取各类球员的数据（整个生涯，单个赛季等）进行数据的清洗和数据缺失的填充。

数据经过简单的预处理后便可进行数据上传操作。使用 ./bin/hdfs dfs -put 本地文件路径命令将本地文件上传到HDFS中，数据格式如图3。

图3 数据格式

3.3 聚类分析

对NBA球员综合能力的评价是广大球迷非常关心的一件事，联盟利用各种各样的奖项来评判球员的素质，比如常规赛MVP，得分王，助攻王，最佳防守球员等，但这些指标都不能体现一个球员的综合素质。单类数据无法对一名球员是否优秀做出综合的合理判断。

球员综合素质的刻画方法选择了Kmeans作为聚类算法，有以下三点原因：1）Kmeans是一种无监督算法，不需要对数据集进行标注。2）Kmeans作为一种迭代式机器学习算法，能充分发挥Spark的性能。3）便于数据可视化。

1. 特征选择：

由于一个球员的助攻数多少很大程度上不能反映出球员组织能力的强弱，所以我们在预处理过的数据中定义了一个新的属性助攻失误比(Assist Turnover Ratio):

ATR = Assist / Turnover

由于失误作为分母不能为零，因此，我们首先将失误为０的行丢弃，这也符合常理，因为打过一整个赛季的球员不太可能没有失误。

最后我们选取了得分(PTS)、篮板(TRB)、助攻失误比(ATR)三个维度对球员进行聚类。

1. 构建ML应用：

我们在本项目中使用的是Spark ML pipeline，提供了一组统一的高级API，它们构建在 DataFrame之上，可帮助用户创建和调整实用的机器学习pipeline。

Pipeline由一系列有顺序的阶段指定，包括转换器和估计器。每个状态的运行是有顺序的，输入的数据框通过每个阶段进行改变。在转换器阶段，transform()方法用于将一个Dataframe转化为另一个Dataframe,并加入新的向量。对于估计器阶段，fit()方法被调用来产生一个转换器,用于作为下一个阶段的输入。

ML应用的关键代码由四部分组成：

1）transform：添加一个助攻失误比的列

nba = nba.withColumn('ATR', (col('AST') / col('TOV')))

2）transform：选择特征向量

vecAssembler = VectorAssembler(inputCols=["TRB", "PTS", "ATR"], outputCol="features")

new\_df = vecAssembler.transform(nba)

3）fit：聚类

kmeans = KMeans(k=5, seed=1)

model = kmeans.fit(new\_df.select('features'))

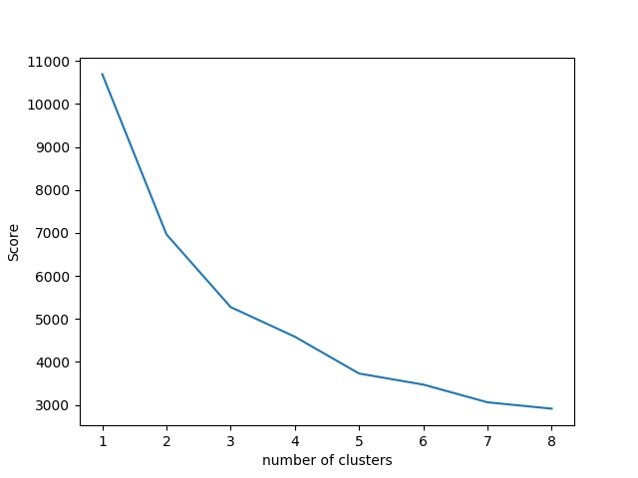
4）transform：可视化

transformed = model.transform(new\_df)

1. 寻找最优K值：

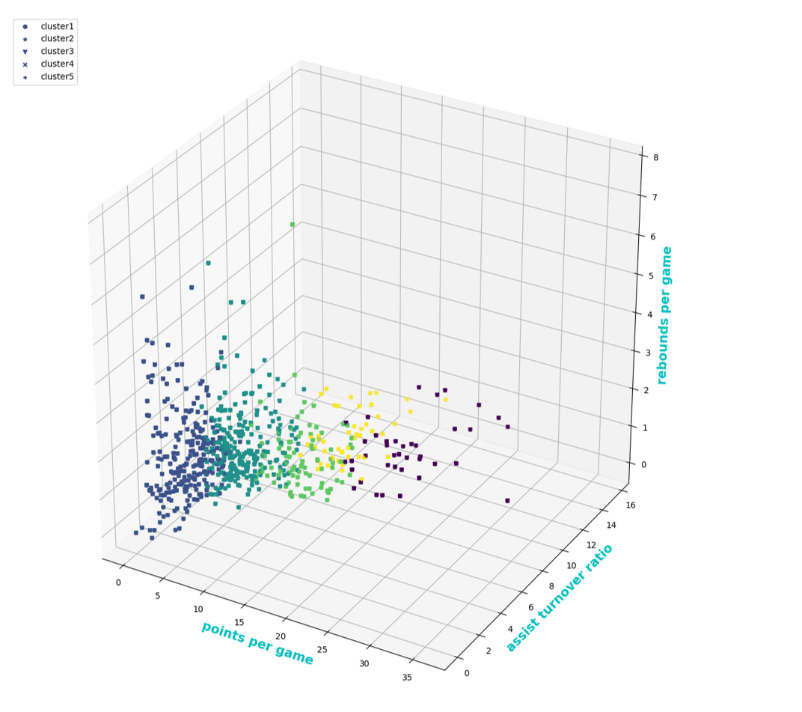
k的选择一般是按照实际需求进行决定，其概念是基于 SSE（sum of the squared errors，误差平方和）作为指标，去计算每一个群中的每一个点，到群中心的距离。且：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数

在我们的实验中得到簇数量与SSE的关系如图4所示，可以看到,这些球员真实类别数应该在3到5之间，为了得到更加细粒度的结果，我们选择使用5作为我们的聚类数。

图5 SSE随簇数量变化曲线

（4）聚类结果可视化：

我们将所有NBA现役球员的数据进行统计，将得分(PTS)、篮板(TRB)、助攻失误比将其数据放在一个三维坐标中展示。得到图5的结果。

图5 聚类结果可视化

图中x轴为场均得分，y轴为场均助攻失误比，z轴为场均篮板，每一个点代表一名球员，不同的颜色代表不同的球员类别。

可以看到最终NBA球员被分为五种类型。我们可以得出以下结论:

1） 一个簇中的球员的水平相当，如果一个簇中有一个球员得到的关注很大，那么该簇内其它球员同样可能得到很大的关注，因为他们属于一个高水平的球员簇中。

2）聚集在右侧的球员是比较优秀的球星，而且数量非常少，这也比较符合我们的认知。

## 4 小结

本案例探讨了大数据背景下的NBA球员分析，通过网络爬虫获取球员数据信息，并进行数据预处理。对球员进行聚类分析，从而对球员的综合素质进行评判。最后，进行球员数据聚类分析结果的可视化。案例结合了体育大数据的背景，用数据支撑了体育产业决策，案例具有良好的实际意义。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习方法完成任务。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），matplotlib（可视化绘图库）。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

参考内容：数据预处理理论、Kmeans聚类算法

1. 吕云翔,张扬编.Python网络爬虫与数据采集[M].人民邮电出版社,2021.

参考内容：数据采集理论