**实验报告：基于聚类分析的网球比赛回合模式研究**

# 一、研究背景与目的

随着现代网球比赛数据记录技术的不断发展，比赛中每一个回合的详细数据得以被精确捕捉和量化。然而，如何从这些海量数据中提取有价值的信息、揭示回合的内在模式，是数据挖掘领域亟待解决的问题之一。本研究旨在通过无监督聚类方法对回合级别的数据进行系统分析，识别回合的特征模式，揭示不同回合的动态特征，并对其进行科学分类。这一研究成果有助于：

**1.剖析选手的回合表现**，揭示击球次数、跑动强度和发球效率等因素对回合模式的影响；

**2.提供量化依据**，为教练团队优化比赛策略提供数据支持；

**3.预测比赛趋势**，识别特定回合模式与获胜概率之间的潜在关联。

# 二、数据概述与预处理

## 2.1 数据来源

数据来源于温布尔登网球锦标赛的详细比赛记录，数据覆盖了多场比赛的每一个回合，包括选手表现、击球次数、发球速度等指标。数据的详细字段涵盖了回合结构、选手技术表现以及关键事件特征。

## 2.2 数据字段描述

数据集共包含 **46** 个字段，每行代表一条比赛回合的数据记录

|  |  |
| --- | --- |
| 变量名 | 变量解释 |
| match\_id | 比赛标识 |
| player1 | 第一位球员的名字（名和姓） |
| player2 | 第二位球员的名字（名和姓） |
| elapsed\_time | 从第一分开始到当前分开始的时间（小时:分钟:秒） |
| set\_no | 比赛中的盘号 |
| game\_no | 盘中的局号 |
| point\_no | 局中的分数号 |
| p1\_sets | 第一位球员赢得的盘数 |
| p2\_sets | 第二位球员赢得的盘数 |
| p1\_games | 第一位球员在当前盘赢得的局数 |
| p2\_games | 第二位球员在当前盘赢得的局数 |
| p1\_score | 第一位球员在当前局中的得分 |
| p2\_score | 第二位球员在当前局中的得分 |
| server | 本分的发球员 |
| serve\_no | 第一发球或第二发球 |
| point\_victor | 本分的胜者 |
| p1\_points\_won | 第一位球员在比赛中赢得的分数 |
| p2\_points\_won | 第二位球员在比赛中赢得的分数 |
| game\_victor | 本分球员赢得了一局 |
| set\_victor | 本分球员赢得了一盘 |
| p1\_ace | 第一位球员发出了无法接触的制胜发球 |
| p2\_ace | 第二位球员发出了无法接触的制胜发球 |
| p1\_winner | 第一位球员击出了无法接触的制胜球 |
| p2\_winner | 第二位球员击出了无法接触的制胜球 |
| winner\_shot\_type | 无法接触的击球类别 |
| p1\_double\_fault | 第一位球员错失了两个发球并输了这一分 |
| p2\_double\_fault | 第二位球员错失了两个发球并输了这一分 |
| p1\_unf\_err | 第一位球员犯了非受迫性失误 |
| p2\_unf\_err | 第二位球员犯了非受迫性失误 |
| p1\_net\_pt | 第一位球员冲到网前 |
| p2\_net\_pt | 第二位球员冲到网前 |
| p1\_net\_pt\_won | 第一位球员在网前赢得了这一分 |
| p2\_net\_pt\_won | 第二位球员在网前赢得了这一分 |
| p1\_break\_pt | 第一位球员有机会在第二位球员发球时赢得一局 |
| p2\_break\_pt | 第二位球员有机会在第一位球员发球时赢得一局 |
| p1\_break\_pt\_won | 第一位球员在第二位球员发球时赢得了这一局 |
| p2\_break\_pt\_won | 第二位球员在第一位球员发球时赢得了这一局 |
| p1\_break\_pt\_missed | 第一位球员错失了在第二位球员发球时赢得一局的机会 |
| p2\_break\_pt\_missed | 第二位球员错失了在第一位球员发球时赢得一局的机会 |
| p1\_distance\_run | 第一位球员在本分中跑动的距离（米） |
| p2\_distance\_run | 第二位球员在本分中跑动的距离（米） |
| rally\_count | 本分中的击球次数 |
| speed\_mph | 发球速度（英里每小时） |
| serve\_width | 发球方向 |
| serve\_depth | 发球的深度 |
| return\_depth | 回球的深度 |

## 2.3 数据预处理

### 2.3.1缺失值填补与类别特征编码

* + speed\_mph 缺失值：分组均值填补（基于set\_no、game\_no）；全局均值填补剩余缺失值。
  + serve\_depth, serve\_width, return\_depth：类别型特征进行编码，并用零值填补缺失值。

1. **特征标准化**：
   * 使用 StandardScaler 对数值型变量进行标准化处理，保证各特征量纲一致。

**三、特征工程**

在数据预处理的基础上，通过特征工程进一步构建了回合级别的统计和动态特征，具体如下：

**3.1 跑动与击球特征**

|  |  |  |
| --- | --- | --- |
| 特征名称 | 定义公式 | 含义说明 |
| 平均跑动距离 (average\_distance\_per\_rally) |  | 每次击球平均的跑动距离 |
| 跑动距离差异 (distance\_diff) | ( | p1\_distance\_run - p2\_distance\_run | ) | 双方选手跑动距离的差异 |

**3.2 技术效率特征**

|  |  |  |
| --- | --- | --- |
| 特征名称 | 定义公式 | 含义说明 |
| ACE效率 (ace\_efficiency) |  | ACE球数与击球次数的比率 |
| 非受迫失误与ACE比率 (unf\_err\_to\_ace\_ratio) |  | 非受迫失误数与ACE球数的比率 |

**3.3 位置与关键回合特征**

|  |  |  |
| --- | --- | --- |
| 特征名称 | 定义公式 | 含义说明 |
| 发球深度方向比率 (serve\_depth\_width\_ratio) |  | 发球深度与方向之间的比率 |
| 回球深度归一化 (return\_depth\_normalized) |  | 回球深度相对于发球深度的归一化 |
| 关键回合标记 (is\_key\_point) | 若回合中出现 ACE 或非受迫失误，则标记为关键回合。 | 二元分类变量，表示回合的重要程度 |

**四、研究方法**

1. **数据准备与标准化**
   * 对构建的特征矩阵进行标准化处理，确保不同尺度的数据可用于聚类分析。
2. **KMeans 聚类分析**
   * 通过肘部法则确定最佳簇数，并应用 KMeans 算法对数据进行聚类。
3. **性能评价指标**
   * 轮廓系数（Silhouette Score）：衡量簇内的紧密性和簇间的分离性。
   * 戴维森堡丁指数（Davies-Bouldin Index）：衡量聚类结果的分离程度。
4. **结果可视化**
   * 使用 PCA 降维至二维和三维空间，对聚类结果进行可视化分析。

**五、实验结果与分析**

**5.1 聚类性能评估**

|  |  |
| --- | --- |
| 指标名称 | 数值 |
| 最优簇数 | 3 |
| 轮廓系数 (Silhouette Score) | 0.4730 |
| 戴维森堡丁指数 (Davies-Bouldin Index) | 0.9756 |

**5.2 簇的特征均值分析**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 簇编号 | 平均击球次数 | 平均发球速度 | 平均跑动距离 | 跑动差异 | ACE效率 | 非受迫失误/ACE比率 | 关键回合比例 |
| 0 | 4.28 | 108.41 | 8.92 | 3.60 | 0 | 0.91 | 1 |
| 1 | 0.95 | 121.03 | 1.68 | 0.43 | 0.52 | 0 | 1 |
| 2 | 2.92 | 112.99 | 7.91 | 3.18 | 0 | 0 | 0 |

**簇 0 分析：复杂回合** 簇 0 具有较高的击球次数和跑动强度，平均击球次数为 4.28，平均跑动距离为 8.92 米。这类回合表现出明显的动态特征，双方选手跑动距离的差异较大（3.60 米），说明选手在这一类回合中频繁移动，并且对抗激烈。此外，非受迫失误与 ACE 球比率较高，关键回合比例为 1，表明这类回合往往在比赛中起到关键作用，是双方选手拼抢激烈的回合。

**簇 1 分析：快速制胜回合** 簇 1 的发球速度最高，平均达到 121.03 英里/小时，但击球次数最低（0.95），平均跑动距离也非常小（1.68 米）。这表明簇 1 的回合往往通过快速发球直接获得分数，ACE 效率达到 0.52，表明选手依赖发球优势快速终结比赛。这类回合的关键回合比例同样为 1，凸显出发球在比赛中的重要性。

**簇 2 分析：常规回合** 簇 2 具有适中的击球次数（2.92）和平均跑动距离（7.91 米），表现为平衡性回合。跑动距离差异较小（3.18 米），ACE 效率和非受迫失误/ACE 比率均接近 0，说明选手在这类回合中主要依靠多拍击球与对手周旋。这类回合的关键回合比例为 0，表明其对比赛的影响较为中性，是比赛中最常见的回合类型。

**5.3 可视化分析**

* **二维可视化**：PCA 组件 1 和 2 的分布图表明，三个簇之间存在显著分离性。
* **三维可视化**：增加 PCA 组件 3，进一步揭示了回合模式在空间中的聚集与分布。

**六、结论与展望**

**6.1 主要结论**

1. 簇 0：击球次数较高，跑动强度大，代表高复杂度回合。
2. 簇 1：发球速度显著高，击球次数少，代表快速制胜回合。
3. 簇 2：击球次数适中，跑动距离平衡，代表常规回合。

**6.2 改进方向**

1. 引入时间序列特征，研究回合的演变动态。
2. 扩展数据规模，验证模型的泛化能力。
3. 深度分析回合模式与比赛胜率的潜在关系，为选手提供战术指导。