Phân tích gom cụm cầu thủ bằng K-means trong dự đoán giá trị chuyển nhượng (ETV)

Người làm: Nguyễn Hải Nam

Ngày: 3 tháng 5 năm 2025

Tài liệu này phân tích số lượng cụm tối ưu để phân loại cầu thủ bóng đá bằng thuật toán K-means, lý do chọn số cụm, và nhận xét về kết quả, sử dụng dữ liệu từ tệp result.csv.

Mục lục

1. Phân tích số lượng cụm tối ưu

Thuật toán K-means được sử dụng để gom các cầu thủ trong tệp result.csv thành các cụm dựa trên các đặc trưng thống kê (như Key Passes, Gls, Save%). Phương pháp Elbow Method được áp dụng để xác định số cụm tối ưu, với kết quả cho thấy số cụm tối ưu là k = 4.

1.1. Phương pháp Elbow Method

Phương pháp Elbow Method phân tích giá trị inertia (tổng bình phương khoảng cách từ các điểm dữ liệu đến tâm cụm gần nhất) qua các giá trị k (số cụm) từ 1 đến 15:

$$\text{Inertia} = \sum_{i=1}^{m} \min_{j} \|x_i - \mu_j\|^2$$

trong đó x_i là điểm dữ liệu, μ_j là tâm cụm, và m là số cầu thủ. Điểm khuỷu tay (elbow point) là giá trị k mà tại đó inertia giảm chậm lại, được xác định bằng KneeLocator.

1.1.1. Kết quả

Biểu đồ Elbow Method (lưu tại histograms/K-means/The optimal number of clusters.png) cho thấy:

- Khi $k = 1, 2$, inertia rất lớn, cho thấy gom cụm kém.
- Từ $k = 1$ đến $k = 4$, inertia giảm mạnh, nghĩa là chất lượng gom cụm cải thiện đáng kể.
- Sau $k = 4$, đường cong inertia trở nên phẳng hơn, cho thấy việc thêm cụm chỉ giảm inertia không đáng kể.

Do đó, KneeLocator xác định k = 4 là điểm khuỷu tay, là số cụm tối ưu.

1.1.2. Lý do chọn k = 4

- Hiệu quả gom cụm: Từ $k = 1$ đến $k = 4$, mỗi cụm thêm vào cải thiện đáng kể chất lượng gom cụm, thể hiện qua sự giảm mạnh của inertia.
- Cân bằng độ phức tạp: Chọn $k = 4$ tránh tạo quá nhiều cụm (dẫn đến quá khớp) hoặc quá ít cụm (không đủ chi tiết).
- Ý nghĩa bóng đá: Số cụm $k = 4$ có thể đại diện cho các nhóm cầu thủ với hiệu suất khác nhau, ví dụ:
  - Cụm 1: Cầu thủ ngôi sao (cao về Gls, Key Passes, Save%).

- Cụm 2: Cầu thủ trung bình khá (hiệu suất ổn định).
- Cụm 3: Cầu thủ trung bình yếu (hiệu suất thấp).
- Cụm 4: Cầu thủ ít thi đấu (thống kê thấp).

1.2. Quy trình thực hiện K-means

Quy trình gom cụm bao gồm:

- Tiền xử lý dữ liệu:
  - Loại bỏ các cột không tính toán (Player, Nation, Team, Position).
  - Chuyển đổi sang kiểu số, điền NaN bằng 0.
  - Chuẩn hóa bằng StandardScaler để trung bình = 0, độ lệch chuẩn = 1.
- Chạy K-means: Thử nghiệm với $k = 1$ đến $k = 15$, sử dụng random_state=42 và n_init=10 để đảm bảo kết quả tái lập.
- Xác định k tối ưu: Sử dụng KneeLocator để tìm điểm khuỷu tay tại $k = 4$.
- Trực quan hóa: Vẽ biểu đồ Elbow Method và lưu tại thư mục histograms/K-means.

2. Kết luận

Số cụm tối ưu để phân loại cầu thủ là k = 4, được xác định bởi phương pháp Elbow Method với điểm khuỷu tay tại $k = 4$. Lựa chọn này cân bằng giữa hiệu quả gom cụm và ý nghĩa bóng đá, phản ánh các nhóm cầu thủ với hiệu suất khác nhau. Tuy nhiên, hạn chế về xử lý NaN, giả định của K-means, và việc không phân tách vị trí cần được cải thiện bằng cách chạy K-means riêng cho từng vị trí, sử dụng phương pháp điền NaN tốt hơn, và thử nghiệm các thuật toán gom cụm khác.