Universidad Nacional de Colombia - Sede Manizales

Informe Trabajo Final: Implementación del Algoritmo K-means y sus Variantes

Santiago Escárraga Vallejo[1], Thomas Giraldo Aguirre[2], Vanessa Osorio Agudelo[3], Daniel Felipe Castellanos Arias[4]

Objetivo

Desarrollar un proyecto en equipo para implementar el algoritmo K-means desde cero, utilizando distintas métricas de distancia. Este ejercicio tiene como finalidad fomentar el trabajo colaborativo y el desarrollo de habilidades en análisis de datos y programación.

Análisis exploratorio

En este trabajo realizamos un análisis exploratorio utilizando los datos proporcionados por los desarrolladores del juego FC25. El objetivo principal es examinar cómo se relacionan entre sí las estadísticas de todos los jugadores del juego. Para ello, empleamos varias herramientas visuales que nos permiten comprender mejor la distribución y la relación de los datos.

Primero, utilizamos un histograma para observar la distribución de los jugadores en función de su (OVR). Esto nos ofrece una forma más clara de ver la cantidad de jugadores según su nivel general. Además, utilizamos diagramas de cajas, los cuales son especialmente útiles para visualizar la distribución del OVR en relación con las diferentes posiciones de campo, incluyendo porteros.

Luego, empleamos una matriz triangular de correlación, que nos permite evidenciar cómo se relaciona cada una de las estadísticas individuales con las demás. Esto facilita la identificación de relaciones significativas entre las distintas variables del juego. Finalmente, en el análisis exploratorio, creamos varios diagramas de cajas adicionales para explorar la correlación de cada estadística desglosada por posición. Esto nos ayuda a observar las diferencias en la relación de las estadísticas según la posición del jugador en el campo.

Implementación del algoritmo K-means

Para la implementación del algoritmo K-means, utilizamos la librería `scikit-learn`, para la clasificación de las posiciones del campo en el videojuego FC25, ya que contamos con una matriz de 45 dimensiones que sería muy compleja de graficar. Por esta razón, utilizamos PCA para la reducción de esta a tres dimensiones, las cuales usaremos más adelante para graficar.

Este proceso comienza definiendo la cantidad de *clusters* que vamos a tener, en nuestro caso cuatro, para cada posición principal del campo de juego. También calculamos los centroides para obtener el valor promedio para un jugador en cualquiera de las cuatro posiciones. Todo esto se logra con el método de clasificación K-means.

[1] sescarraga@unal.edu.co
[2] thgiraldoa@unal.edu.co
[3] vaosorioa@unal.edu.co
[4] dacastellanosar@unal.edu.co

Posteriormente, a cada jugador se le asigna un centroide para encontrar el punto medio para una de las cuatro posiciones. Este proceso se repite hasta que la distancia entre el centroide y el jugador sea menor a un valor ya establecido en el código, lo que nos permite clasificar a los jugadores según sus posiciones. Finalmente, para facilitar la visualización, creamos una gráfica 2D y 3D con toda la información explicada.

Implementación del algoritmo K-means++

El algoritmo K-means++ es una mejora del algoritmo K-means clásico. Su objetivo principal es la agrupación de los datos mediante la mejora en la selección inicial de los centroides. En el presente trabajo se realizó la implementación de dicho algoritmo, el cual consta de los siguientes pasos:

- Inicializar el proceso escogiendo al azar un solo centroide, tomando un punto aleatorio del conjunto de datos.
- Calcular la distancia entre todos los demás puntos y el centroide inicial.
- Cuando ya se tiene más de un centroide, calcular la distancia de cada punto al centroide más cercano.
- Seleccionar el siguiente centroide basado en la probabilidad de los puntos que estén más alejados del centroide más cercano.

Este proceso se repite hasta que se hayan seleccionado los k centroides necesarios, garantizando que el nuevo centroide esté bien separado del anterior y que los grupos sean más inteligentemente distribuidos.

Implementación del algoritmo K-means con distancia euclidiana

El algoritmo comienza eligiendo un conjunto de jugadores al azar utilizando la librería `random`, a los cuales se les asignan datos promedio iniciales basados en las posiciones del campo, conocidos como centroides.

La distancia euclidiana se utiliza para asignar a cada jugador el centroide más cercano, según la fórmula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

donde x_1, x_2, y_1, y_2 representan las estadísticas del jugador y del centroide en un espacio bidimensional.

Este proceso se repite iterativamente hasta que las posiciones de los centroides se estabilizan.

Implementación del algoritmo K-means con distancia Mahalanobis

La distancia de Mahalanobis es una métrica que mide la distancia entre un punto y una distribución multivariada. Utilizaremos la fórmula:

$$d_M(x, c) = \sqrt{(x - c)^T S^{-1} (x - c)}$$

donde x es un vector de 45 dimensiones, c es el centroide y S^{-1} es la inversa de la matriz de covarianza. Calculamos esta distancia usando la función `mahalanobis` de la librería SciPy.

Implementación del algoritmo K-means con distancia L1 (Manhattan)

El algoritmo comienza eligiendo un grupo de jugadores al azar, y luego selecciona un número de *clusters*. La distancia Manhattan, dada por la fórmula:

$$|x_1 - x_2| + |y_1 - y_2|$$

se utiliza para asignar a cada jugador al centroide más cercano. Este proceso se repite hasta que los centroides se estabilizan.

Conclusiones

El desarrollo del proyecto permitió al equipo analizar y comprender los fundamentos para la implementación de los algoritmos K-means y K-means++, además de las variaciones con distancia euclidiana, Mahalanobis y Manhattan, fomentando el trabajo en equipo y la colaboración.

Referencias

- Kaggle: Davis Nyagamid. (2024). EA Sports FC 25 Database Ratings and Stats. *Kaggle*. Recuperado de

  `https://www.kaggle.com/datasets/nyagami/ea-sports-fc-25-database-ratings-and-stats`

- Caminos Aleatorios. (2020, abril 16). Tan cerca y tan lejos: La distancia de Mahalanobis. Recuperado de:

  `https://caminosaleatorios.wordpress.com/2020/04/16/`

- Esri. (2024). Comprender el análisis de distancia euclidiana. Recuperado de

  `https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-analyst/.htm`

- Jain, N. (2023, May 26). Mastering data clustering: Your comprehensive guide to K-means and K-means++. *AI Accelerator Institute*. Recuperado de `https://www.aiacceleratorinstitute.com/mastering-data-clustering-your-comprehensive-guide-to-k-means-and-k-means/`

- LinkedIn. (2023). ¿Cómo comparas la distancia de Mahalanobis? Recuperado de

  `https://www.linkedin.com/advice/0/how-do-you-compare-mahalanobis-distance?lang=esoriginalSubdomain=es`

.