Aplicación de PCA y K-means para la reducción de dimensionalidad y clustering en un conjunto de datos balanceado

Bustamante Callisaya Hansel Alain

Introducción

La minería de datos y el análisis de grandes volúmenes de información han crecido considerablemente en los últimos años. Con la disponibilidad de bases de datos más grandes y complejas, se hace necesario utilizar técnicas que nos permitan procesar y analizar estos datos de manera eficiente. En este artículo, se exploran dos técnicas populares para la reducción de dimensionalidad y la segmentación de datos: **Principal Component Analysis (PCA)** y **K-means clustering**.

El objetivo de este trabajo es analizar cómo PCA se puede usar para reducir las dimensiones de un conjunto de datos sin perder información clave y cómo K-means clustering puede ayudar a identificar patrones o agrupamientos en los datos. Para lograr esto, se aplicarán estas técnicas a un conjunto de datos balanceados de fútbol, con el propósito de clasificar jugadores según sus características.

1. Reducción de dimensionalidad con PCA

1.1. ¿Qué es PCA?

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos mientras se conserva la mayor variabilidad posible de las características originales. Este proceso es crucial en muchos problemas de aprendizaje automático, especialmente cuando los datos contienen muchas características (dimensiones), lo que puede llevar a un sobreajuste (overfitting) y aumentar la complejidad computacional.

En términos sencillos, PCA transforma los datos en un nuevo sistema de coordenadas donde las nuevas características, llamadas **componentes principales**, son combinaciones lineales de las características originales, pero ordenadas por la cantidad de varianzas que explican.

1.2. Aplicación de PCA en el conjunto de datos

En este trabajo, se utiliza PCA para reducir la dimensionalidad de un conjunto de datos balanceados de fútbol. El conjunto de datos contiene diversas características numéricas de los jugadores, y el objetivo es reducir el número de dimensiones mientras se conserva la mayor cantidad de información posible.

Para lograr esto, primero se cargan los datos y se separan las características (X) de la etiqueta de clasificación (y). Luego, se realiza una transformación PCA para proyectar los datos en un espacio de menor dimensión.

```
# Cargar el dataset balanceado
df_resampled = pd.read_csv('proy_balanceado.csv')

# Seleccionar características (X) y etiqueta (y)
X = df_resampled.drop('categoria_jugador', axis=1)
y = df_resampled['categoria_jugador']

# Inicializar PCA con el número deseado de componentes
principales
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
```

En este caso, se seleccionan 2 componentes principales para la visualización, lo que permite graficar los datos en un espacio bidimensional.

1.3. Resultados de la reducción de dimensionalidad con PCA

Al aplicar PCA, los datos originales se proyectan sobre dos componentes principales, lo que reduce la complejidad sin perder demasiada información. La siguiente gráfica muestra cómo se distribuyen los datos después de la reducción de dimensionalidad:

```
plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.5, c='blue',
label="Datos")
plt.title('Reducción de dimensionalidad con PCA (2
Componentes)')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.show()
```

2. Aplicación de K-means para clustering

2.1. ¿Qué es el clustering?

El clustering es una técnica de aprendizaje no supervisado utilizada para agrupar elementos similares en clusters o grupos. Uno de los algoritmos de clustering más conocidos y utilizados es **K-means**, que agrupa los datos en k clusters de acuerdo con su proximidad en el espacio de características.

2.2. Aplicación de K-means en el conjunto de datos

Una vez que hemos reducido la dimensionalidad con PCA, podemos aplicar **K-means** para identificar posibles agrupamientos dentro de los datos. En este caso, se realiza un clustering con tres grupos (clusters):

```
# Inicializar K-means con 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X)

# Obtener los centroides y las etiquetas de los clusters
centroids = kmeans.cluster_centers_
labels = kmeans.labels_
```

2.3. Visualización de los resultados de clustering

Para visualizar los resultados del clustering, se grafican los datos reducidos por PCA, coloreados según las etiquetas asignadas por K-means. Además, se muestran los centroides de los clusters en la gráfica.

```
plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=labels,
cmap='viridis', alpha=0.5)
plt.scatter(centroids[:, 0], centroids[:, 1], s=300, c='red',
marker='X', label="Centroides")
plt.title('Clusters identificados por K-means')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.show()
```

Los resultados muestran cómo K-means ha agrupado los jugadores en tres clusters distintos, lo que puede reflejar distintas categorías de jugadores basadas en sus características.

3. Evaluación de los resultados y conclusiones

3.1. Resultados obtenidos

En el proceso de reducción de dimensionalidad con PCA y posterior clustering con K-means, se lograron dos objetivos importantes:

1. Reducir la dimensionalidad de los datos sin perder demasiada variabilidad.
2. Identificar agrupamientos dentro de los jugadores que podrían representar diferentes tipos o categorías de jugadores.

Las gráficas obtenidas muestran que el uso de PCA permitió representar los datos de manera eficiente en 2 dimensiones, mientras que K-means identificó tres clusters distintos, lo cual podría indicar diferentes categorías o perfiles de jugadores.

3.2. Discusión

Aunque PCA y K-means son técnicas poderosas, es importante destacar que la elección del número de componentes principales y el número de clusters puede influir en los resultados. En este caso, se seleccionaron arbitrariamente 2 componentes principales para la visualización y 3 clusters, pero estas elecciones deben ser validadas dependiendo del conjunto de datos y el problema en cuestión.

3.3. Conclusión

El uso de PCA y K-means en el conjunto de datos de fútbol permitió obtener una representación más sencilla de los datos y descubrir patrones subyacentes en los jugadores. Sin embargo, es crucial ajustar los parámetros de estas técnicas según las características específicas de los datos para obtener los mejores resultados.