UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

Artículo

Análisis de Datos con PCA y K-Means Clustering

Responsable: Badani Dávalos Kassandra Andrea

Asignatura: Inteligencia Artificial (INF-354)

Fecha: 9 de diciembre

La Paz – Bolivia

2024

Análisis de Datos con PCA y K-Means Clustering: Reducción de Dimensionalidad y Agrupamiento de Datos en un Conjunto de Canciones

Introducción

En el contexto de la analítica de datos y el aprendizaje automático, el procesamiento y análisis de grandes volúmenes de información es crucial para obtener conclusiones significativas. En este artículo, aplicamos dos técnicas fundamentales para la reducción de la complejidad en grandes conjuntos de datos: **Principal Component Analysis (PCA)** y **K-Means Clustering**. Estas técnicas no solo mejoran la eficiencia computacional, sino que también permiten descubrir patrones y estructuras subyacentes en los datos. Utilizamos un conjunto de datos de canciones, cuyo objetivo es analizar sus características musicales y asociar patrones con su popularidad.

1. Principal Component Analysis (PCA)

El **Análisis de Componentes Principales (PCA)** es una herramienta matemática poderosa que permite transformar un conjunto de datos de características correlacionadas en un conjunto de componentes principales no correlacionados. Esta transformación ayuda a reducir la dimensionalidad de los datos, preservando la mayor cantidad posible de variabilidad, lo que facilita la visualización, el análisis y mejora la eficiencia computacional de los modelos.

1.1 Preparación de los Datos

El primer paso en el análisis fue la creación de una nueva variable binaria denominada **Is_Popular**, que clasifica las canciones en populares o no populares. Para ello, calculamos el cuartil 75% de la columna **Points (Total)**, y las canciones con una puntuación mayor o igual a este valor fueron clasificadas como populares.

El siguiente paso consistió en separar el conjunto de datos en variables predictoras (**X**) y la variable objetivo (**y**). Se utilizó un **80/20 split** para entrenar y probar los modelos, respectivamente, asegurando que el modelo tuviera suficiente representación de las canciones populares y no populares.

1.2 Aplicación de PCA

Para aplicar PCA, probamos diferentes cantidades de componentes principales (3, 5, 9, 10, 11, 12), con el objetivo de identificar la cantidad mínima de componentes que aún conservaran suficiente información del conjunto original de características. Se usó **Random Forest** para la clasificación, evaluando los modelos con métricas como la **precisión**, **recall**, **F1-score** y **RMSE**.

El modelo con un mayor número de componentes (12) y el de menor dimensionalidad (5) ofrecieron resultados muy similares en términos de precisión (94%) y **RMSE** de

aproximadamente 0.24. Este hallazgo sugirió que la reducción de dimensionalidad mediante PCA no afectó significativamente el rendimiento, y que un número menor de componentes podía ser suficiente para mantener la capacidad predictiva del modelo.

1.3 Resultados de PCA

Los resultados obtenidos en la evaluación mostraron que la reducción de dimensionalidad mediante PCA no solo mejoró la interpretabilidad de los datos, sino que también permitió obtener modelos más rápidos sin perder rendimiento. Con 12, 10, 9 y 5 componentes principales, el **accuracy** se mantuvo alrededor del 94%, mientras que el **RMSE** no variaba significativamente, lo que implica que las predicciones eran consistentes. Este comportamiento destacó la eficiencia de PCA en la retención de la variabilidad esencial, a la vez que simplificaba los datos, lo que puede reducir el tiempo de cómputo en entornos con grandes volúmenes de datos.

El rendimiento constante a lo largo de diferentes cantidades de componentes sugiere que los datos tienen una estructura simple, donde una pequeña cantidad de dimensiones (o componentes principales) es suficiente para explicar la mayor parte de la variabilidad presente en las características musicales de las canciones.

2. K-Means Clustering

El **K-Means Clustering** es un algoritmo de agrupamiento no supervisado que agrupa datos en clústeres (o grupos) basados en sus similitudes. El objetivo del K-Means es minimizar la distancia entre los puntos de datos dentro de un clúster y el centroide de ese clúster. Este enfoque es útil cuando no se tiene información previa sobre las etiquetas de los datos, y es particularmente relevante para la segmentación de canciones en grupos con características musicales similares.

2.1 Preparación y Escalado de Datos

Para aplicar K-Means, primero se escaló el conjunto de datos utilizando **StandardScaler**. El escalado de las características fue esencial, ya que K-Means es sensible a la escala de los datos, y características con diferentes escalas (como *loudness* vs. *danceability*) podrían dominar el algoritmo. Al estandarizar los datos, todas las características tienen la misma media (0) y desviación estándar (1), lo que garantiza que ninguna variable influya desproporcionadamente en el agrupamiento.

2.2 Método del Codo para Seleccionar el Número de Clústeres

Una vez escalados los datos, se aplicó el **Método del Codo** para determinar el número óptimo de clústeres. El método del codo se basa en analizar la **WCSS** (Within-Cluster Sum of Squares), que mide la variabilidad dentro de cada clúster. El número de clústeres que minimiza esta variabilidad, sin seguir disminuyendo de manera significativa, se considera el óptimo.

En nuestro caso, el análisis mostró que el número óptimo de clústeres se encontraba entre 3 y 5. Después de realizar pruebas con diferentes valores de **k**, se seleccionaron 5 clústeres para el análisis, dado que la disminución de la **WCSS** mostraba un "codo" claro en el gráfico en ese punto.

2.3 Resultados de K-Means

Tras aplicar K-Means con 5 clústeres, se observaron patrones claros en las características musicales que definieron cada grupo. A continuación, se describen las principales características de cada uno de los clústeres:

- **Clúster 0**: Canciones con niveles moderados de *energy* y alta *danceability*. Este grupo parece representar canciones aptas para bailar, con un ritmo pegajoso y una energía que las hace atractivas para audiencias en eventos sociales.
- **Clúster 1**: Canciones más tranquilas, con bajos niveles de *energy* y *loudness*. Estas canciones, con una menor *valence*, parecen ser más suaves, probablemente representando géneros como baladas o música ambiental.
- **Clúster 2**: Canciones con *valence* positiva elevada y alta *energy*, lo que sugiere un estilo más optimista y vibrante, típico de géneros como el pop y el dance.
- **Clúster 3**: Canciones con baja *danceability* y *energy*, lo que podría indicar canciones con un ritmo menos accesible o más experimental, como música alternativa o de nicho.
- **Clúster 4**: Canciones de energía moderada y *loudness* media, predominantemente de la región de Oceanía. Este clúster parece representar un estilo musical relajado y más melódico, con influencias regionales.

Además, se observó que un número considerable de canciones provenía de **Latinoamérica**, especialmente en los clústeres 0 y 2, lo que podría indicar una tendencia hacia un estilo musical más optimista y rítmico en esta región.

3. Discusión de los Resultados

3.1 Interpretación de PCA

La reducción de dimensionalidad mediante **PCA** permitió identificar las principales características que contribuyen a la variabilidad en los datos de canciones. Aunque el modelo fue capaz de mantener un alto nivel de precisión con solo unos pocos componentes, el uso de PCA no solo simplificó el modelo, sino que también ayudó a mitigar el sobreajuste al reducir el ruido inherente a las características menos relevantes. En resumen, PCA ayudó a mejorar la eficiencia computacional sin sacrificar la calidad predictiva del modelo.

3.2 Interpretación de K-Means

El algoritmo **K-Means** permitió descubrir patrones claros en el comportamiento musical de las canciones, segmentándolas en clústeres con características similares. Esta segmentación no solo ofrece una mejor comprensión de los diferentes estilos musicales presentes en el conjunto de datos, sino que también puede ser útil para la toma de decisiones en áreas como marketing musical, recomendaciones personalizadas y análisis de tendencias musicales por regiones.

Por ejemplo, la observación de que el clúster 4 está predominantemente compuesto por canciones de Oceanía sugiere que ciertos géneros pueden estar asociados con regiones geográficas específicas, lo que puede ser aprovechado por las plataformas de streaming para personalizar las recomendaciones en función de la ubicación del usuario.

Conclusión

El análisis realizado utilizando **PCA** y **K-Means Clustering** ha demostrado ser eficaz para reducir la dimensionalidad de los datos, mejorar la eficiencia computacional y descubrir patrones significativos en el conjunto de datos de canciones. La combinación de estas técnicas facilita el análisis de grandes volúmenes de datos musicales y proporciona una comprensión más profunda de las características que definen la popularidad y los estilos musicales. Estos métodos tienen aplicaciones potenciales en áreas como marketing musical, análisis de tendencias y recomendación de música personalizada.

Referencias

1. **Jolliffe, I. T.** (2002). *Principal

Component Analysis* (2nd ed.). Springer Series in Statistics. ISBN: 978-0387954424. 2. **Lloyd, S. P.** (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory*, 28(2), 129–137. DOI: 10.1109/TIT.1982.1056489. 3. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. ISBN: 978-0387848570. 4. **Bishop, C. M.** (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN: 978-0387310732. 5. **Duda, R. O., Hart, P. E., & Stork, D. G.** (2001). *Pattern Classification* (2nd ed.). Wiley. ISBN: 978-0471056690.