|  |  |
| --- | --- |
|  |  |
| Proyecto No Supervisado. Técnicas de aprendizaje de máquina. |  |
| Pontificia Universidad Javeriana - Wikipedia |  |
|  | 27/05/2024Entregables. |
|  | Juan Esteban Rivera y Victoria Chavarro |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | *Abstract*En este proyecto se aborda el problema de agrupar usuarios en función de sus calificaciones de películas, utilizando técnicas de aprendizaje no supervisado. Se aplicaron tres enfoques principales: K-means, Agglomerative Clustering y Agglomerative Clustering combinado con la Factorización de Matrices No Negativas (NMF). El análisis incluyó la normalización robusta de los datos y la reducción de dimensionalidad mediante PCA y NMF. Los resultados mostraron que Agglomerative Clustering con NMF produjo clusters más compactos y homogéneos, aunque con tamaños desiguales. Este enfoque capturó mejor los patrones latentes en las calificaciones de los usuarios, destacando la importancia de NMF en la mejora del clustering. Las distancias intra-cluster fueron significativamente menores en comparación con K-means, y las distancias entre clusters indicaron una buena separación. Estos hallazgos sugieren que NMF es una herramienta valiosa para identificar grupos de usuarios con preferencias similares de manera más precisa y detallada.*Introducción* *La personalización en los sistemas de recomendación ha adquirido una relevancia significativa en la era digital, donde la cantidad de contenido disponible puede abrumar a los usuarios. Las plataformas de streaming, comercio electrónico y redes sociales dependen en gran medida de sistemas de recomendación eficaces para mejorar la experiencia del usuario al proporcionar contenido relevante. En este contexto, las técnicas de aprendizaje no supervisado, como el clustering, juegan un papel crucial en la segmentación de usuarios basándose en sus comportamientos y preferencias.*  *El objetivo de este proyecto es agrupar a los usuarios en función de sus calificaciones de películas utilizando diferentes técnicas de clustering. La agrupación efectiva de usuarios permite a los sistemas de recomendación personalizar las sugerencias de manera más precisa y mejorar la satisfacción del usuario.*  *El problema abordado es cómo identificar y agrupar a los usuarios que tienen preferencias similares en un gran conjunto de datos de calificaciones de películas. La complejidad del problema radica en la alta dimensionalidad y la presencia de valores atípicos en los datos, lo que puede afectar la precisión y la utilidad de los clusters formados.*  *Para abordar este problema, se utilizaron varias técnicas de aprendizaje no supervisado, incluidas K-means, Agglomerative Clustering y Agglomerative Clustering combinado con la Factorización de Matrices No Negativas (NMF). La implementación de estas técnicas se realizó en múltiples pasos, comenzando con la normalización de los datos para reducir la influencia de los valores atípicos, seguida por la aplicación de algoritmos de clustering para formar los grupos de usuarios.* *Aproximación del problema* *Para abordar la pregunta de cómo agrupar a los usuarios basándonos en sus calificaciones de películas, se emplearon las siguientes técnicas de aprendizaje no supervisado:*  *1.* ***Normalización de Datos:***  *- StandardScaler: Inicialmente, los datos fueron normalizados utilizando `StandardScaler` para transformar las características a una escala con media 0 y desviación estándar 1. Sin embargo, para manejar mejor los valores atípicos, se utilizó `RobustScaler`, que utiliza la mediana y el rango intercuartílico (IQR) para escalar los datos, reduciendo la influencia de los valores extremos.*  *2.* ***Reducción de Dimensionalidad:***  *- PCA (Análisis de Componentes Principales): Se aplicó PCA para reducir la dimensionalidad de los datos y facilitar la visualización en 2D, lo que ayudó a entender la distribución de los usuarios en el espacio reducido.*  *- NMF (Factorización de Matrices No Negativas): Se utilizó NMF para descomponer la matriz de calificaciones en representaciones latentes de usuarios y películas, permitiendo capturar patrones latentes en las preferencias de los usuarios.*  *3.* ***Algoritmos de Clustering:***  *- K-means: Este método se aplicó para formar clusters iniciales. K-means requiere especificar el número de clusters de antemano y utiliza la media de los puntos dentro de un cluster como centroide.*  *- Agglomerative Clustering: Se utilizó para crear una jerarquía de clusters mediante un enfoque ascendente, comenzando con cada punto de datos como un cluster individual y fusionando iterativamente los clusters más similares.*  *- Agglomerative Clustering con NMF: Combinando NMF con Agglomerative Clustering, se buscó mejorar la compactación de los clusters al utilizar representaciones latentes obtenidas mediante NMF, que capturan mejor los patrones subyacentes en las calificaciones de los usuarios.*  *El proceso comenzó con la preparación y normalización de los datos, seguida por la aplicación de PCA y NMF para reducir la dimensionalidad. Posteriormente, se aplicaron los algoritmos de clustering a los datos transformados. Se evaluaron los resultados mediante la visualización de los clusters y el cálculo de distancias intra-cluster y entre clusters, lo que permitió comparar la efectividad de cada enfoque.*  *La combinación de NMF y Agglomerative Clustering resultó en clusters más compactos y homogéneos, aunque con tamaños desiguales, destacando la capacidad de NMF para mejorar la precisión del clustering al manejar mejor la estructura subyacente de los datos. Este enfoque permitió identificar grupos de usuarios con preferencias similares de manera más precisa, mejorando potencialmente la efectividad de los sistemas de recomendación basados en estos clusters.* *Resultados* *Al aplicar las técnicas de aprendizaje no supervisado al problema de agrupación de usuarios basados en sus calificaciones de películas, se obtuvieron los siguientes resultados:*  ***K-means***  *1. Distribución de Clusters:*  *- Se formaron cinco clusters con tamaños variables. Dos clusters resultaron vacíos, mientras que los otros tres clusters presentaron una significativa dispersión interna.*  *- La distribución desigual y la alta dispersión en algunos clusters indicaron que K-means no manejó bien la variabilidad y los valores atípicos en los datos.*  *2. Distancias Intra-Cluster y Entre Clusters:*  *- Las distancias intra-cluster promedio fueron mayores, especialmente para uno de los clusters con una distancia intra-cluster de 95.65.*  *- Las distancias entre clusters fueron grandes, lo que sugiere una buena separación entre clusters, pero a costa de una mayor dispersión interna.*  ***Agglomerative Clustering***  *1. Distribución de Clusters:*  *- Se formaron cuatro clusters con tamaños más equilibrados y una mejor compactación en comparación con K-means.*  *- Los clusters mostraron una estructura más clara y homogénea, con menor dispersión interna.*  *2. Distancias Intra-Cluster y Entre Clusters:*  *- Las distancias intra-cluster promedio fueron menores en comparación con K-means, indicando clusters más compactos.*  *- Las distancias entre clusters fueron moderadas, proporcionando una buena separación sin comprometer la homogeneidad dentro de los clusters.*  ***Agglomerative Clustering con NMF***  *1. Distribución de Clusters:*  *- Se formaron cuatro clusters con tamaños desiguales. El Cluster 0 fue significativamente más grande que los otros clusters.*  *- A pesar de la desigualdad en los tamaños, los clusters fueron más homogéneos y compactos en términos de calificaciones.*  *2. Distancias Intra-Cluster y Entre Clusters:*  *- Las distancias intra-cluster promedio fueron las más bajas entre los tres enfoques, con valores de 0.39, 1.67, 0.79 y 0.00, respectivamente, indicando una excelente compactación.*  *- Las distancias entre clusters fueron menores que las observadas en K-means, pero comparables a las de Agglomerative Clustering sin NMF, mostrando una buena separación.*  *3. Resumen de Calificaciones:*  *- Cluster 0: Promedio de calificaciones de 3.54 con una distribución uniforme y una mediana de 4.0.*  *- Cluster 1: Promedio de calificaciones de 3.19 con alta concentración en calificaciones de 3 y 4 estrellas, y una mediana de 3.0.*  *- Cluster 2: Promedio de calificaciones de 3.48 con una mediana de 3.5.*  *- Cluster 3: Promedio de calificaciones de 3.40 con una mediana de 3.5.*  ***Conclusiones de los Resultados***  *- Compactación y Homogeneidad: La combinación de NMF y Agglomerative Clustering resultó en clusters significativamente más compactos y homogéneos, destacando la efectividad de NMF para capturar patrones latentes en los datos.*  *- Separación de Clusters: Aunque los clusters obtenidos con NMF y Agglomerative Clustering fueron de tamaños desiguales, ofrecieron una buena separación y mejor representación de las preferencias de los usuarios.*  *- Mejora con NMF: NMF ayudó a mejorar la precisión del clustering al reducir la influencia de valores atípicos y variabilidad excesiva, lo que resultó en una mejor agrupación de usuarios con preferencias similares.*  *Estos resultados sugieren que la integración de NMF con técnicas de clustering jerárquico puede proporcionar una agrupación más precisa y útil para aplicaciones de sistemas de recomendación, permitiendo una mejor personalización y satisfacción del usuario.*  ***Caracterización de los clusters para el negocio***  *Cluster 0:*  *Películas Destacadas:*  *• Forrest Gump (1994)*  *• Shawshank Redemption, The (1994)*  *• Pulp Fiction (1994)*  *• Silence of the Lambs, The (1991)*  *• Matrix, The (1999)*  *Características:*  *• Calificación Promedio: Alta.*  *• Año Promedio: Principalmente de los 90s.*  *• Géneros Predominantes: Drama, Crimen, Ciencia Ficción.*  *Descripción: Este cluster agrupa a usuarios que tienen una preferencia por películas icónicas de los años 90s, con un fuerte énfasis en dramas y thrillers que han sido ampliamente aclamados por la crítica y el público.*  *Cluster 1:*  *Películas Destacadas:*  *• 2001: A Space Odyssey (1968)*  *• 28 Days Later (2002)*  *• 300 (2007)*  *• A.I. Artificial Intelligence (2001)*  *• Ace Ventura: Pet Detective (1994)*  *Características:*  *• Calificación Promedio: Varía, pero generalmente positiva.*  *• Año Promedio: Mezcla de películas desde finales de los 60s hasta principios de los 2000s.*  *• Géneros Predominantes: Ciencia Ficción, Acción, Comedia.*  *Descripción: Este cluster agrupa a usuarios que disfrutan de una mezcla de ciencia ficción clásica, acción moderna y comedia, indicando una amplia variedad de gustos cinematográficos.*  *Cluster 2:*  *Películas Destacadas:*  *• Dark Knight, The (2008)*  *• Inception (2010)*  *• Matrix, The (1999)*  *• Star Wars: Episode IV - A New Hope (1977)*  *• Up (2009)*  *Características:*  *• Calificación Promedio: Alta.*  *• Año Promedio: Principalmente de finales de los 90s y 2000s.*  *• Géneros Predominantes: Acción, Aventura, Ciencia Ficción, Animación.*  *Descripción: Los usuarios en este cluster tienen una fuerte afinidad por películas de acción y ciencia ficción modernas, así como por películas animadas de alta calidad. Este grupo incluye tanto clásicos modernos como favoritos recientes.*  *Cluster 3:*  *Películas Destacadas:*  *• ’burbs, The (1989)*  *• ’night Mother (1986)*  *• …And Justice for All (1979)*  *• 10 Things I Hate About You (1999)*  *• 101 Dalmatians (1996)*  *Características:*  *• Calificación Promedio: Moderada.*  *• Año Promedio: Desde los 70s hasta finales de los 90s.*  *• Géneros Predominantes: Comedia, Drama, Familia.*  *Descripción: Este cluster agrupa a usuarios con una inclinación hacia películas de comedia y drama de varias décadas, así como películas familiares. Muestra una mezcla de clásicos de los 70s y 80s con comedias adolescentes de los 90s* | |  |