IDENTIFICACIÓN DE PATRONES EN BASES DE DATOS DE PELÍCULAS USANDO CLUSTERING

Rubén Esquivias[1]
1 Universidad Autónoma de Nuevo León, Facultad de Ciencias Físico Matemáticas, San Nicolás de los Garza, Nuevo León, 664554, México

RESUMEN

Pendiente....

Key Words: Pendientes.....

1. INTRODUCCIÓN

En este trabajo se busca analizar un conjunto de datos de películas partiendo de una base de datos de **Agregar referencia*** con el propósito de identificar agrupamientos naturales entre ellas, es decir, ver si existen "familias" de películas con características similares. Para ello se emplean técnicas de aprendizaje no supervisado, particularmente el método de k-means, que permite dividir el conjunto de datos en grupos con base en la similitud entre sus variables.

Este tipo de análisis puede llegar a tener múltiples aplicaciones prácticas y brindar conocimientos en este caso para comprender la estadística de cada dato recopilado por los usuarios. Por ejemplo, puede ayudar a plataformas de streaming a recomendar contenidos de forma más precisa, o servir como apoyo para estudios de mercado que buscan entender qué tipo de películas suelen tener más éxito entre distintos públicos.

2. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos utilizado contiene información general sobre diversas películas, incluyendo variables como el título, el año de lanzamiento, la duración, la puntuación promedio otorgada por los usuarios, la cantidad total de votos y una medida de popularidad. Antes de comenzar el análisis, fue necesario realizar una limpieza del dataset para eliminar valores faltantes y asegurar que todas las columnas relevantes tuvieran un formato numérico adecuado.

Además, las variables fueron normalizadas, de modo que cada una tuviera el mismo peso en el proceso de agrupamiento. Esto es importante porque una variable con valores mucho más grandes (como el número de votos) podría dominar el cálculo de distancias y alterar el resultado final. Una vez preparado el conjunto, se realizó un análisis exploratorio para observar las tendencias generales de los datos y tener una idea previa de qué relaciones podrían existir entre las variables.

	Unnamed: 0	id	title	overview	release_date	popularity	vote_average	vote_count
0	0	278	The Shawshank Redemption	Imprisoned in the 1940s for the double murder ...	9/23/1994	26.9579	8.712	28675
1	1	238	The Godfather	Spanning the years 1945 to 1955, a chronicle o...	3/14/1972	26.5804	8.686	21701
2	2	240	The Godfather Part II	In the continuing saga of the Corleone crime f...	12/20/1974	15.6559	8.571	13099
3	3	424	Schindler's List	The true story of how businessman Oskar Schind...	12/15/1993	12.5642	8.565	16616
4	4	389	12 Angry Men	The defense and the prosecution have rested an...	4/10/1957	14.6028	8.549	9307
...	...	...	...	...	...	...	...	...
8555	8555	238603	Earth to Echo	After a construction project begins digging in...	6/14/2014	1.6137	5.900	593
8556	8556	11968	Into the Blue	When they take some friends on an extreme spor...	9/30/2005	3.3432	5.902	1539

Fig. 1.

3. ANTECEDENTES

Revisar el artículo: Understanding the confluence of retailer characteristics, market characteristics and online pricing strategies que habla sobre el modelo aplicado

así como: ***Insertar referencia de los datos*** para visualizar los datos

4. METODOLOGIA

Para este estudio se estructuraron distintas fases con el propósito de identificar patrones, estructuras y algunas cuantas relaciones predictivas posibles dentro de una base de datos de películas. El análisis se centró en variables numéricas como **calificación**, **popularidad**, **presupuesto** y **duración**, entre otras, buscando agrupar y modelar el comportamiento de las cintas mediante técnicas de *machine learning* supervisadas y no supervisadas.

Preprocesamiento de datos

Primero, se realizó una limpieza general del conjunto de datos eliminando valores nulos y registros incompletos. Posteriormente, se aplicó una **normalización mediante estandarización**, con el fin de

evitar que variables con escalas distintas influyeran de forma desigual en los algoritmos.

La estandarización se realiza con la siguiente fórmula:

$$z = \frac{x - \mu}{\sigma}$$

donde: -

$$(x)$$

= valor original de la variable -

$$(\mu)$$

= media de la variable -

$$(\sigma)$$

= desviación estándar

De esta manera, todas las variables quedaron con media 0 y desviación estándar 1, lo que garantiza la comparabilidad entre ellas.

Agrupamiento con DBSCAN

Utilizando el algoritmo **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**, que forma grupos a partir de regiones densamente pobladas, separando las zonas de baja densidad como ruido se realiza el siguiente proceso.

Un punto p pertenece a un clúter si existe al menos un punto q tal que:

$$dist(p, q) \leq \varepsilon$$

y q tiene al menos *minsamples* vecinos dentro de ese radio. Los puntos que no cumplen esta condición se etiquetan como ruido ($cluster = -1$).

Este enfoque resulta muy bueno para la base de datos de películas, ya que la distribución no sigue una forma geométrica regular y el algoritmo no requiere especificar el número de grupos como tal.

Tras esto, se exploraron distintas combinaciones de los parámetros ** (eps)** y **minsamples** para obtener la mejor configuración posible. La calidad de cada uno de los los agrupamientos se evaluó mediante el **índice de Davies–Bouldin (DBI)**:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{s_i + s_j}{d_{ij}} \right)$$

donde:

$$(s_i) :$$

dispersión promedio de los puntos del clúster (i)

$$(d_{ij}) :$$

distancia entre los centroides de los clústeres

$$(i)$$

y

$$(j)$$

$$(k) :$$

número total de clústeres

Cuanto menor es el DBI, mejor es la separación y cohesión entre los grupos.

Modelos supervisados

Con el fin de extender el análisis y realizar predicciones sobre las características de las películas (por ejemplo, su calificación o nivel de popularidad), se incorporaron algoritmos de aprendizaje **supervisado** los cuales son: **Regresión Lineal** y **Random Forest**. Se parte a desglozar cada uno de estos.

Regresión Lineal

El modelo busca encontrar una relación lineal entre las variables predictoras

$$(x_i)$$

y la variable respuesta

$$(y)$$

:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

donde:

$$(\hat{y}) :$$

valor estimado de la variable dependiente

$$(\beta_i) :$$

coeficientes ajustados del modelo

$$(\epsilon) :$$

término de error

Este modelo se utilizó para aproximar tendencias generales y evaluar cómo ciertos atributos influyen en el rendimiento o aceptación de las películas.

Random Forest

El algoritmo **Random Forest** está basado en el principio de *ensembles*, combinando múltiples árboles de decisión entrenados con subconjuntos aleatorios de los datos. Cada árbol genera una predicción, y el resultado final se obtiene mediante el **promedio** (para regresión) o **voto mayoritario** (para clasificación).

Matemáticamente, el modelo se expresa como:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

donde

$$(f_t(x))$$

es la predicción del árbol

$$(t)$$

y

$$(T)$$

es el número total de árboles en el bosque. Este método reduce el sobreajuste (*overfitting*) y mejora la estabilidad de las predicciones.

Evaluación de desempeño

Para evaluar los modelos supervisados, se consideraron métricas comunes de error como el **MAE (Mean Absolute Error)** y el **RMSE (Root Mean Square Error)**, definidos respectivamente como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Estas métricas cuantifican la diferencia promedio entre los valores reales y los predichos, permitiendo medir la precisión de los modelos utilizados. Primero se exploraron distintos valores de k (el número de clusters), con el fin de identificar cuántos grupos representaban de forma más coherente la estructura de los datos. Luego se ejecutó el algoritmo y se graficaron los resultados para visualizar cómo se distribuían las películas dentro de cada cluster. Finalmente, se realizó un conteo por grupo para conocer qué tan equilibrado era el reparto entre ellos.

También algunos algoritmos de tipo supervisados que se agregaron fueron el Random Forest y la Regresión de tipo Lineal. Esto ya que maneja relaciones no lineales y reduce el sobreajuste gracias al ensamble de los datos. Posteriormente se describirá la parte matemática.

5. RESULTADOS

Se mejoró la gráfica de visualización de los clusters haciendo que los puntos de interés se concentren en toda la región visible.

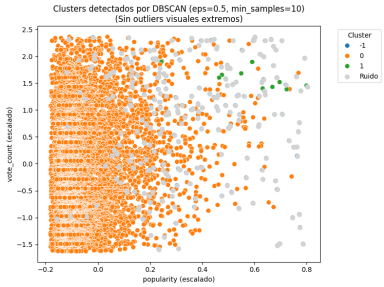

Fig. 2.

Asimismo se procedió a realizar los modelos matemáticos combinando los subconjuntos aleatorios de datos y variables. Usando específicamente Random Forest y Regresión lineal.

Ambos modelos presentan errores bajos; **Random Forest** logra un mejor desempeño global. Las métricas obtenidas confirman que los datos presentan relaciones no lineales. Se observaron altas correlaciones entre $vote_count$ y $vote_average$, coherentes con el comportamiento esperado.

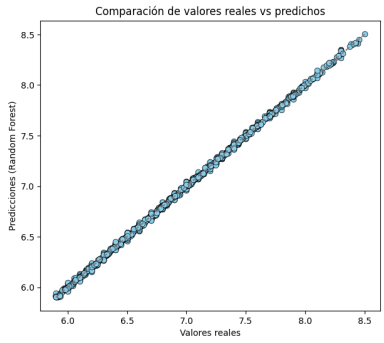

Fig. 3.

6. CONCLUSIONES Y DISCUSIÓN

De momento con las gráficas obtenidas se encuentran algunas limitantes tal y como que hay falta de variables de texto (‘overview‘) y categóricas (‘género‘), las cuales podrían mejorar el desempeño.

De cierto modo la regresión lineal ofrece interpretabilidad, mientras que Random Forest captura mejor las no linealidades. Se puede incorporar el procesamiento de texto (NLP), extracción de año desde $release_date$, y validación cruzada para afinar algunos hiperparámetros.

Finalmente, los resultados del agrupamiento fueron representados gráficamente en un espacio bidimensional mediante diagramas de dispersión,

mostrando la distribución de las películas en función de variables relevantes como la popularidad y el número de votos. Asimismo, se realizó un conteo de instancias por grupo para evaluar la homogeneidad de los clústeres obtenidos y la proporción de ruido detectado.

Esta combinación de enfoques —no supervisado y supervisado— permitió no solo identificar patrones naturales dentro del conjunto de películas, sino también explorar su capacidad predictiva en función de los atributos más influyentes.

REFERENCIAS

[1] Delen, D., Sharda, R., Kumar, P. (2007). Movie forecast guru: A Web-based DSS for Hollywood managers. Decision Support Systems, 43(4), 1151–1170. https://doi.org/10.1016/j.dss.2006.03.012

[2] Hastie, Tibshirani Friedman (2009). The Elements of Statistical Learning.

[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1).

[4] James, Witten, Hastie Tibshirani (2021). An Introduction to Statistical Learning.

- Documentación de *scikit-learn*: https://scikit-learn.org/stable/