Facultad de Matemática y Computación. Universidad de La Habana.
Sistemas de Recuperación de Información.

Hibridación de técnicas de Sistemas de Recomendación. Ventajas del enfoque probabilístico en comparación con la factorización matricial. Implementación del algoritmo Naive Bayes Collaborative Filtering y su expansión para realizar recomendaciones a grupos de usuarios.

Daniel Machado Pérez - daniel.machado.0206@gmail.com
Daniel Toledo Martínez - daniel020126@gmail.com
Osvaldo R. Moreno Prieto - osvaldo0202013@gmail.com
August 31, 2024

Resumen

El presente trabajo explora la hibridación de técnicas en sistemas de recomendación, centrándose en las ventajas del enfoque probabilístico frente a la factorización matricial en cuanto a filtrado colaborativo. Se destaca cómo el enfoque probabilístico, al proporcionar una representación explícita de las incertidumbres, mejora la interpretabilidad y explicación de las recomendaciones generadas. En particular, se implementa el algoritmo *Naive Bayes Collaborative Filtering* (NBCF), que combina la simplicidad del modelo *Naive Bayes* con el poder del filtrado colaborativo, permitiendo recomendaciones precisas y explicativas. Además, se expande este algoritmo para adaptarse a la recomendación a grupos de usuarios, abordando un área clave en la personalización colectiva de contenidos. Los resultados demuestran que el enfoque probabilístico no solo ofrece una alternativa robusta a la factorización matricial, sino que también potencia la capacidad del sistema para ofrecer recomendaciones personalizadas y comprensibles, tanto a individuos como a grupos.

Palabras Clave: Sistemas de Recomendación (RS), Filtrado Colaborativo (CF), Enfoque Probabilístico, Factorización Matricial, *Naive Bayes Collaborative Filtering* (NBCF), Recomendación para Grupos.

1 Introducción

1.1 Descripción del tema y Técnicas de Recomendación

Los sistemas de recomendación se han consolidado como herramientas esenciales en la personalización de contenidos en diversas plataformas digitales, desde servicios de *streaming* hasta comercio electrónico. Estos sistemas tienen como objetivo filtrar grandes volúmenes de información y presentar a los usuarios elementos relevantes según sus preferencias. Entre las técnicas de recomendación más utilizadas, se destacan el filtrado colaborativo, el filtrado basado en contenido, el filtrado demográfico y los enfoques híbridos que combinan estos métodos.

El filtrado colaborativo, en particular, ha sido ampliamente adoptado debido a su capacidad para identificar patrones de comportamiento entre usuarios y ofrecer recomendaciones basadas en similitudes en sus interacciones previas. Este enfoque se puede implementar mediante técnicas basadas en memoria, que utilizan directamente las interacciones pasadas de los usuarios, o mediante técnicas basadas en modelos, que crean representaciones abstractas de las relaciones entre usuarios e ítems.

1.2 Enfoques de Filtrado Colaborativo Basado en Modelos

Dentro del filtrado colaborativo basado en modelos, dos enfoques destacan por su eficacia y popularidad: la factorización matricial y los modelos probabilísticos. La factorización matricial, como lo demuestra el algoritmo de descomposición en valores singulares (SVD), es una técnica poderosa para descomponer la matriz de interacciones usuario-ítem en factores latentes, permitiendo predicciones precisas de las preferencias de los usuarios. No obstante, su principal limitación radica en la falta de interpretabilidad de los factores latentes, lo que dificulta la explicación de las recomendaciones generadas.

En contraste, los modelos probabilísticos, como el *Naive Bayes Collaborative Filtering* (NBCF), ofrecen una alternativa que, si bien puede alcanzar niveles de precisión similares a los de la factorización matricial, presenta la ventaja adicional de proporcionar interpretaciones más claras de las recomendaciones. El enfoque probabilístico permite modelar explícitamente la incertidumbre en las preferencias de los usuarios, lo que facilita la explicación del porqué de cada recomendación.

1.3 Antecedentes y Justificación

La elección del enfoque probabilístico como base de esta investigación se sustenta en los hallazgos presentados en la tesis doctoral titulada "Sistema recomendador híbrido basado en modelos probabilísticos". Esta tesis profundiza en las ventajas de utilizar modelos probabilísticos en

sistemas de recomendación, destacando su capacidad para superar las limitaciones de los enfoques tradicionales de factorización matricial. Además, se presenta una implementación del algoritmo NBCF, que ha mostrado resultados prometedores en términos de precisión y explicabilidad.

Sin embargo, un área poco explorada en esta tesis es la capacidad de estos modelos para realizar recomendaciones a grupos de usuarios, una característica esencial en contextos como la recomendación de contenido para familias, grupos de amigos o equipos de trabajo. Esta investigación se propone expandir el algoritmo NBCF, siguiendo las recomendaciones de la tesis doctoral, para adaptarlo a la recomendación grupal, un desafío significativo en la personalización colectiva.

!!!!!!!!!!!!!!!!!! EXPLICAR SOLUCION NUESTRA !!!!!!!!!!!

1.4 Dataset Seleccionado

Para la evaluación de la implementación y expansión del algoritmo NBCF, se ha seleccionado el dataset MovieLens[6], un conjunto de datos ampliamente utilizado en la investigación de sistemas de recomendación. MovieLens contiene millones de calificaciones de películas proporcionadas por usuarios, lo que lo convierte en un recurso valioso para el análisis y desarrollo de modelos de recomendación. La riqueza y diversidad del dataset permiten probar la eficacia de los algoritmos en un entorno cercano a escenarios del mundo real. Este dataset fue uno de los utilizados en la tesis antes mencionada.

1.5 Estructura del Trabajo

El presente informe se estructura en varias secciones que desarrollan en detalle los diferentes aspectos de la investigación:

- **Estado del Arte**: Se revisa la literatura existente sobre técnicas de recomendación, con un enfoque en el filtrado colaborativo basado en modelos.
- **Algoritmo NBCF**: Se describe la implementación del algoritmo NBCF y su funcionamiento.
- **Expansión del Algoritmo NBCF**: Se presenta la adaptación del NBCF para realizar recomendaciones a grupos de usuarios, detallando las modificaciones realizadas.
- **Evaluación de los Resultados**: Se analizan los resultados obtenidos tras la implementación y se comparan con enfoques tradicionales.
- **Conclusiones**: Se resumen los hallazgos más relevantes de la investigación y se sugieren posibles direcciones futuras.

2 Estado del Arte

Los sistemas de recomendación se han vuelto indispensables en la era de la información, donde los usuarios requieren herramientas que les permitan descubrir contenidos relevantes de manera eficiente. Existen diversas técnicas para abordar este problema, cada una con sus propias ventajas y limitaciones. Entre las más destacadas están el filtrado colaborativo, el filtrado basado en contenido, el filtrado demográfico y los enfoques híbridos. En esta sección, se revisarán las principales técnicas de recomendación, con un enfoque particular en el filtrado colaborativo y sus variantes basadas en modelos probabilísticos.

2.1 Técnicas de Recomendación

- **Filtrado Colaborativo**: Este enfoque se basa en la idea de que los usuarios que han compartido preferencias similares en el pasado probablemente coincidan en sus elecciones futuras. El filtrado colaborativo puede implementarse a través de dos métodos: basado en memoria y basado en modelos. Los enfoques basados en memoria, como el algoritmo de k vecinos más cercanos (k-NN), utilizan directamente la matriz de interacciones usuario-ítem para realizar recomendaciones. Por otro lado, los enfoques basados en modelos, que incluyen técnicas como la factorización matricial y los modelos probabilísticos, construyen un modelo predictivo a partir de los datos disponibles, ofreciendo recomendaciones más precisas y escalables. [2]
- **Filtrado Basado en Contenido**: Este método recomienda ítems a un usuario en función de la similitud entre los ítems que ha consumido previamente y otros ítems disponibles. A diferencia del filtrado colaborativo, se basa en las características de los ítems, como el género, el director o los actores en el caso de películas. [2]
- **Filtrado Demográfico**: Aunque menos utilizado en comparación con los métodos anteriores, el filtrado demográfico se basa en las características personales de los usuarios, tales como su edad, género o ubicación. Este enfoque supone que usuarios con características demográficas similares tienden a compartir preferencias similares. Si bien puede ser útil para ciertos contextos, su efectividad suele ser menor, ya que no tiene en cuenta las interacciones individuales entre usuarios e ítems. [2]
- **Enfoques Híbridos**: Estos combinan dos o más de las técnicas mencionadas para mejorar la precisión y superar las limitaciones inherentes a cada uno de los métodos. Por ejemplo, un sistema híbrido puede combinar el filtrado colaborativo con el filtrado basado en contenido para ofrecer recomendaciones más completas, tanto en precisión como en diversidad. [2]

2.2 Filtrado Colaborativo Basado en Modelos

El filtrado colaborativo basado en modelos ha demostrado ser especialmente eficaz en sistemas de recomendación a gran escala. Entre los enfoques más destacados se encuentran la factorización matricial y los modelos probabilísticos.

- **Factorización Matricial**: Esta técnica ha demostrado ser una de las más efectivas para el filtrado colaborativo. En la factorización matricial, la matriz de interacciones usuario-ítem se descompone en dos matrices de menor dimensión que representan factores latentes tanto para los usuarios como para los ítems. Estos factores latentes permiten realizar predicciones sobre las preferencias de los usuarios al capturar características no observadas explícitamente. Aunque la factorización matricial, especialmente con algoritmos como la descomposición en valores singulares (SVD), ha demostrado ser muy precisa, su principal limitación radica en la falta de interpretabilidad. Los factores latentes no siempre son comprensibles o intuitivos para los usuarios, lo que dificulta la explicación de las recomendaciones. [4]

- **Modelos Probabilísticos**: En contraste con la factorización matricial, los modelos probabilísticos proporcionan una representación más clara de las incertidumbres en las preferencias de los usuarios. Uno de los enfoques más representativos es el *Naive Bayes Collaborative Filtering* (NBCF), que combina la simplicidad del modelo de *Naive Bayes* con la estructura del filtrado colaborativo. Este enfoque permite una mayor interpretabilidad, ya que ofrece una explicación probabilística de las recomendaciones. Además, el NBCF ha mostrado ser altamente adaptable a diferentes escenarios, permitiendo la incorporación de nuevas variables sin comprometer su eficiencia. [4]

2.3 Desarrollo en las Tesis y Papers

El enfoque probabilístico ha sido objeto de un estudio detallado en la tesis doctoral titulada "Sistema recomendador híbrido basado en modelos probabilísticos"[4]. En esta tesis, se aborda la integración de modelos probabilísticos dentro de sistemas de recomendación híbridos, destacando cómo estos modelos no solo permiten una mayor precisión, sino que también aportan una capa de interpretabilidad que los métodos de factorización matricial no ofrecen. El autor propone un enfoque híbrido que combina los beneficios del filtrado colaborativo basado en modelos probabilísticos con técnicas de filtrado basado en contenido.
Otra tesis doctoral relevante es la titulada "Recomendación a grupos de usuarios usando el concepto de singularidades"[5]. En este trabajo, se introduce un enfoque innovador para realizar recomendaciones a grupos de usuarios, considerando las preferencias individuales dentro del grupo, pero ajustando las recomendaciones para maximizar la satisfacción colectiva. Este tipo de investigación es crucial para la expansión del algoritmo NBCF, que se centra en recomendaciones individuales, hacia escenarios donde las recomendaciones deben adaptarse a un colectivo de usuarios con preferencias diversas.
Además, en el paper *"A Collaborative Filtering Approach Based on Naive Bayes Classifier"*[1], se profundiza en la implementación del NBCF y se demuestra su viabilidad como alternativa a los métodos tradicionales de filtrado colaborativo. Los resultados obtenidos en este estudio muestran que el NBCF puede igualar o superar el rendimiento de la factorización matricial, especialmente en datasets donde la interpretabilidad es tan importante como la precisión.
Finalmente, el trabajo *"Hybrid Collaborative Filtering Based on Users' Rating Behavior"*[3] presenta un enfoque híbrido que integra el comportamiento de valoración de los usuarios con el filtrado colaborativo. Este enfoque tiene una relevancia particular para nuestro proyecto, ya que permite ajustar las recomendaciones no solo en función de las interacciones pasadas, sino también considerando la manera en que los usuarios valoran los ítems, lo que aporta una capa adicional de personalización.

3 Algoritmo NBCF

3.1 Introducción al Algoritmo NBCF

El algoritmo *Naive Bayes Collaborative Filtering* (NBCF) es una técnica innovadora dentro del campo de los sistemas de recomendación colaborativos. A diferencia de otros enfoques, como la factorización matricial, el NBCF aprovecha la simplicidad y efectividad del clasificador *Naive Bayes* para predecir las preferencias de los usuarios en función de sus interacciones anteriores con ítems. Este método considera la probabilidad de que un usuario asigne una cierta calificación a un ítem, basándose en las calificaciones previas tanto del usuario como de otros usuarios con comportamientos similares.

3.2 Formulación Matemática del Algoritmo NBCF

El algoritmo NBCF se basa en la combinación de dos enfoques principales: basado en usuarios y basado en ítems. En cada uno de estos enfoques, se calcula la probabilidad a priori de que un usuario califique un ítem con un valor específico, y posteriormente se calcula el *likelihood* para ajustar esta probabilidad en función de las calificaciones observadas.

- **Enfoque basado en el usuario**: la probabilidad a priori y el *likelihood* se calculan de acuerdo con los ítems que cada usuario ha votado. [4]

- **Enfoque basado en ítems**: la probabilidad a priori y el *likelihood* se calculan de acuerdo con los votos que cada ítem ha recibido. [4]

- **Enfoque híbrido**: integra los enfoques basados en el usuario e ítems, a fin de complementarse uno con otro y mejorar la precisión del modelo. [4]

Para el desarrollo de cada uno de estos enfoques se utiliza los siguientes conceptos de probabilidades:

- **Probabilidad A Priori**: En el enfoque basado en ítems, se calcula la probabilidad a priori de que un usuario u asigne una calificación y a un ítem i, denotado como $P(r_u = y)$. De manera análoga, en el enfoque basado en usuarios, se calcula la probabilidad de que un ítem i reciba una calificación y de cualquier usuario u, denotado como $P(r_i = y)$.

$$P(r_i = y) = \frac{|\{u \in U | r_{u,i} = y\}| + \alpha}{|\{u \in U | r_{u,i} \neq \bullet\}| + |R| * \alpha} \quad (1)$$

  [4]

  Donde:

  - U es el conjunto de usuarios.
  - $r_{u,i}$ es la calificación otorgada por el usuario u al ítem i.
  - α es un parámetro para evitar 0 probabilidades.
  - $|R|$ representa el número de votos plausibles.
  - $\bullet$ representa la ausencia de voto.

- **Likelihood**: El *likelihood* ajusta la probabilidad a priori mediante la consideración de la información adicional disponible en las calificaciones observadas. Para el enfoque basado en ítems, esto se expresa como $P(r_v = k | r_u = y)$, que representa la probabilidad de que otro usuario v califique con k un ítem que ha sido calificado con y por el usuario u. Similarmente, para el enfoque basado en usuarios, se calcula el *likelihood* correspondiente $P(r_j = k | r_i = y)$.

$$P(r_j = k | r_i = y) = \frac{|\{u \in U | r_{u,j} = k \wedge r_{u,i} = y\}| + \alpha}{|\{u \in U | r_{u,j} \neq \bullet \wedge r_{u,i} = y|\} + |R| * \alpha} \quad (2)$$

  [4]

- **Combinación de Enfoques**: En el enfoque híbrido, se integran las probabilidades obtenidas de los enfoques basados en usuarios y en ítems, proporcionando un modelo más robusto y preciso para la predicción de calificaciones.

  $P(r_{u,i} = y)$ representa el valor de probabilidad de que el usuario u vote el ítem i con el voto y:

$$P(r_{u,i} = y) \propto \left(P(r_u = y) \cdot \prod_{v \in Ui} P(r_v = r_{v,i} | r_u = y) \right)^{\frac{1}{1+|Ui|}} \cdot \left(P(r_i = y) \cdot \prod_{j \in Iu} P(r_j = r_{u,j} | r_i = y) \right)^{\frac{1}{1+|Iu|}} \quad (3)$$

  [4]

  Donde:

  - $I_u = \{i \in I \mid r_{u,i} \neq \bullet\}$ es el conjunto de ítems votados por el usuario u,
  - y $U_i = \{u \in U \mid r_{u,i} \neq \bullet\}$ es el conjunto de usuarios que han votado el ítem i.

3.3 Algoritmo NBCF: Implementación Paso a Paso

El algoritmo NBCF se implementa de manera iterativa, asegurando la eficiencia computacional mediante técnicas de memorización que permiten evitar el recálculo innecesario de probabilidades. A continuación se describen los pasos del algoritmo:

- **Inicialización**: Se inicializan las probabilidades a priori y los contadores utilizados en el cálculo de *likelihoods*.
- **Iteración sobre Usuarios e Ítems**: Para cada usuario, se calcula la probabilidad de cada calificación posible basada en las calificaciones observadas para los ítems que ha evaluado. De manera similar, se calcula para cada ítem la probabilidad de recibir una calificación específica basada en las calificaciones anteriores recibidas.
- **Almacenamiento de Resultados**: Los valores calculados se almacenan para ser utilizados posteriormente en la predicción de nuevas calificaciones, evitando la necesidad de recalcular durante la fase de predicción.

Este enfoque garantiza que el NBCF no solo sea eficiente, sino que también se adapte bien a problemas de gran escala, manteniendo una complejidad computacional similar a la de otros métodos avanzados, como la factorización matricial [5].

3.4 Resultados Experimentales y Comparativa

El algoritmo NBCF ha demostrado su eficacia en múltiples conjuntos de datos públicos (MovieLens, FilmTrust, Yahoo, BookCrossing)[4], superando en varias métricas clave a los métodos de referencia más utilizados en el campo:

- Error Medio Absoluto (MAE)

- Precisión y *Recall*
- Ganancia acumulada descontada normalizada (nDCG)

Se compararon los siguientes enfoques:

- NBCF (usuario)
- NBCF (ítem)
- NBCF (híbrido)
- BNMF
- GGM
- INBM
- Bi-CF
- NMF

Los resultados fueron los siguientes:

- **MovieLens**: El enfoque híbrido de NBCF ha mostrado mejoras significativas en medidas de MAE, precisión y *recall*, así como el enfoque basado en ítems fue mejor en la nDCG en comparación con enfoques tradicionales.[4]
- **FilTrust**: el MAE de NBCF (híbrido) logra mejores resultados que los otros dos enfoques propuestos, mientras que la precisión y recall son mejores con NBCF (ítems) y NBCF (usuario). Por otro lado, cuando aumenta el número de recomendaciones, nDCG es mejor con el enfoque NBCF (híbrido).[4]
- **Yahoo**: nDCG es mejor en NBCF (híbrido) en comparación con NBCF (ítem) y NBCF (usuario). Además, la precisión y el *recall* de los tres enfoques propuestos presentan un resultado casi similar entre ellos. Así mismo, hay una superioridad lograda en MAE de NBCF (híbrido) con respecto a los otros enfoques propuestos.[4]
- **BookCrossing**: NBCF (híbrido) y NBCF (ítem) proveen mejores resultados para nDCG en comparación con los métodos de línea base de CF. A diferencia de otros conjuntos de datos en BookCrossing las métricas de precisión y *recall* son mejores para los métodos GGM, INBM y Bi-CF. Sin embargo muestran una mejora con respecto a los métodos BNMF y NMF. NBCF (híbrido) se muestra superior al resto de los enfoques en cuanto al MAE.[4]

3.5 Conclusión

El algoritmo NBCF representa una mejora significativa en el ámbito de los sistemas de recomendación colaborativos, combinando la simplicidad del clasificador *Naive Bayes* con técnicas de filtrado colaborativo para ofrecer recomendaciones precisas y eficaces. Su capacidad para integrar múltiples enfoques y adaptarse a diferentes escenarios lo convierte en una herramienta valiosa para la mejora de la experiencia del usuario en plataformas de recomendación.

4 Expansión del Algoritmo NBCF

5 Evaluación de los resultados

6 Conclusiones

References

[1] Valdiviezo-Diaz, P., Ortega, F., Cobos, E., & Lara-Cabrera, R. (2019). A collaborative filtering approach based on Naïve Bayes classifier. IEEE Access, 7, 108581-108592.

[2] González, O. E., & Jacques, S. M. (2017). Estado del arte en los sistemas de recomendación. Res. Comput. Sci., 135, 25-40.

[3] Ortega, F., Rojo, D., Valdiviezo-Diaz, P., & Raya, L. (2018). Hybrid collaborative filtering based on users rating behavior. IEEE Access, 6, 69582-69591.

[4] Valdiviezo, P. M. (2019). Sistema recomendador híbrido basado en modelos probabilísticos (Doctoral dissertation, Universidad Politécnica de Madrid).

[5] Ortiz, R. H. (2020). Recomendación a grupos de usuarios usando el concepto de singularidades. (Doctoral dissertation, Universidad Politécnica de Madrid).

[6] Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4), 1-19. https://doi.org/10.1145/2827872