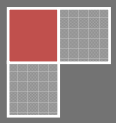
2012

Aprendizaje para Clasificación con Factorización Matricial Basado en Listwise para Filtrado Colaborativo

Iván López Espejo

22/04/2012



Sumario

1. Introducción

Actualmente, los sistemas de recomendación de contenido son un importante punto de atención para los investigadores a causa de la gran cantidad de contenido multimedia disponible en Internet al alcance de los usuarios: libros, música, películas, etc. Particularmente, el **filtrado colaborativo** se ha erigido como una de las técnicas de recomendación más satisfactorias, la cual se basa en la idea de que a un usuario le podrá gustar probablemente un contenido que le gusta a otros usuarios con sus mismas o similares preferencias. El propósito más importante que un sistema de recomendación debe cumplir es el de proveer al usuario de una lista de recomendación de contenido. En base a esta idea, el trabajo que aquí se recoge trata de una extensión escalable a la aproximación de factorización matricial aplicada al filtrado colaborativo denominada ListRank-MF (donde MF viene de *Matrix Factorization*). Esta aproximación hace uso de una técnica de aprendizaje para clasificación listwise con el fin de clasificar u ordenar los diferentes elementos multimedia para cada uno de los usuarios, donde los usuarios y dichos elementos multimedia son representados como características latentes aprendidas haciendo uso de factorización matricial. La contribución principal de la técnica aquí expuesta es doble: por un lado proporciona un rendimiento superior en términos de recomendación sobre el estado del arte basado en la aproximación de la factorización matricial y, por otro, mantiene una complejidad lineal con la cantidad de ratings observados en la matriz usuario-elemento multimedia dada, pudiendo, por tanto, escalarla con el fin de ser usada en colecciones de contenido realmente grandes.

El aprendizaje para clasificación (de ahora en adelante LTR por sus siglas en inglés) es una técnica de machine-learning supervisado que construye automáticamente un modelo de clasificación o ranking a partir de unos datos de entrenamiento. Recientemente, el LTR ha sido objeto de intensivos esfuerzos de investigación, de donde resulta el ejemplo Yahoo! LTR Challenge, pues repercute en beneficios directos sobre las técnicas de recuperación de información (*information retrieval*) mediante el envío de peticiones, y de recomendación a partir de perfiles de usuario (como es el caso del presente trabajo analizado).

2. Estado del Filtrado Colaborativo y del LTR

El filtrado colaborativo puede ser basado en memoria o en modelo. En general, las aproximaciones basadas en memoria hacen recomendaciones partiendo de la base de las similitudes entre usuarios (basadas en usuarios) o entre elementos multimedia (basadas en elementos multimedia). De otro lado, las aproximaciones basadas en modelo en un primer momento ajustan modelos de predicción basados en datos de entrenamiento y luego usan dichos modelos con el fin de predecir las preferencias de los usuarios sobre determinados contenidos multimedia. Además, las técnicas de factorización matricial han atraído la atención de los investigadores debido a las mejoras que proporciona en términos de escalabilidad y precisión, especialmente en entornos ingentes de contenido multimedia. Las técnicas de factorización matricial normalmente aprenden características latentes de los usuarios y los elementos multimedia a partir de los ratings observados en las matrices usuario/elemento multimedia, posteriormente usadas para predecir ratings no observados.

Dentro del área del filtrado colaborativo, la atención en términos de investigación se ha movido desde el problema de la predicción del rating al problema de la calidad en la clasificación o la lista de recomendación que el sistema genera. No obstante, las aproximaciones existentes basadas en lo anterior (como la clasificación probabilística bayesiana) tienen un alto coste de cómputo, lo que limita su escalabilidad. No obstante, ListRank-MF aquí descrito presenta una complejidad lineal con el número de ratings observados dada una matriz de ratings usuario/elemento multimedia.

En cuanto al LTR según una aproximación listwise, un ejemplo de entrenamiento individual es una lista de elementos multimedia completa. Las funciones de pérdida para el LTR listwise se formulan para medir la distancia entre la lista de referencia y la lista de salida del modelo de clasificación. Varios algoritmos son aplicados para aprender el modelo de clasificación óptimo local o global. Se propuso a la probabilidad de permutación para representar la lista de clasificación, la cual podía ser simplificada a la probabilidad de que un elemento multimedia dado sea clasificado en primera posición para una lista dada. Esta última probabilidad es la que emplea ListRank-MF para representar la lista de recomendación, haciendo que dicha técnica se acerque al LTR listwise.

3. Algoritmos

El marco de trabajo de la técnica ListRank-MF está basado en la factorización matricial probabilística, cuyos fundamentos se recogen en [2] y donde una factorización matricial es formulada a partir de inferencia estadística sobre distribuciones condicionales de ratings observados, así como sobre distribuciones a priori de ratings de usuarios y de ratings de elementos multimedia. Dicho marco de trabajo se formula como

$$U,V = \underset{U,V}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}\left(R_{ij} - g\left(U_i^T V_j\right)\right)^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2.$$

La factorización matricial probabilística busca representar la matriz de ratings usuario-elemento multimedia, R, como dos matrices U y V, donde M es el número de usuarios y N es la cantidad de elementos multimedia. Se emplea un conjunto de características latentes d-dimensionales para representar tanto U como V. Los subíndices de dichas matrices indican una determinada columna de ellas compuesta por un vector de características d-dimensional. De otro lado R_{ij} denota el rating del usuario i-ésimo sobre el elemento multimedia j-ésimo. I_{ij} es una función indicadora que es igual a 1 cuando $R_{ij} > 0$ y 0 en otro caso. Por último, λ_U y λ_V son coeficientes de regularización donde, normalmente, $\lambda_U = \lambda_V = \lambda$. La función logística $g\left(U_i^T V_j\right)$ se emplea para acotar el rango de su argumento, siendo

$$g\left(U_i^T V_j\right) = \frac{1}{1 + e^{-U_i^T V_j}}.$$

La probabilidad de que un elemento multimedia dado (el j-ésimo) sea clasificado en primera posición para una lista dada (la del usuario i-ésimo) se puede calcular como

$$P_{l_i}\left(R_{ij}\right) = \frac{\phi\left(R_{ij}\right)}{\sum_{k=1}^{K}\phi\left(R_{ik}\right)},$$

donde se ha supuesto la existencia de K elementos multimedia para el usuario i-ésimo y $\phi(x)$ es usualmente la función exponencial, pues se precisa que $\phi(x)$ sea estrictamente creciente y estrictamente positiva.

El ListRank-MF se formula como la función de pérdida a partir de la entropía cruzada de las anteriores probabilidades de elementos multimedia

en las listas de ejemplo de entrenamiento y en las listas de clasificación a partir del modelo de clasificación:

$$L(U,V) = \sum_{i=1}^{M} \left\{ -\sum_{j=1}^{N} P_{l_i}(R_{ij}) \log P_{l_i}\left(g\left(U_i^T V_j\right)\right) \right\} + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2).$$

Las listas de ejemplo de entrenamiento consisten en conjuntos de elementos multimedia de entrenamiento en los perfiles de cada usuario. La salida del modelo de recomendación es una lista de recomendación para cada uno de los usuarios compuesta de elementos ordenados de forma descendente en función del rating de acuerdo con el valor de $U_i^T V$.

La anterior función de pérdida representa la incertidumbre entre las listas de entrenamiento y las listas de salida del modelo de clasificación. El modelo de clasificación óptimo debería de proveer la mínima incertidumbre entre las listas de ratings de entrenamiento y las listas de predicciones de salida. En este caso tenemos que la factorización matricial está optimizada para posiciones de clasificación de elementos multimedia en las listas de los usuarios.

Finalmente, puesto que la función de pérdida no es convexa conjuntamente sobre U y V, se escoge usar gradiente descendiente fijando alternativamente U y V, a partir de los cuales puede obtenerse un mínimo local. El vector gradiente se obtendría de aplicar derivadas parciales sobre la función de pérdida, de la forma

$$\nabla L(U,V) = \begin{pmatrix} \frac{\partial L(U,V)}{\partial U_i} \\ \frac{\partial L(U,V)}{\partial V_j} \end{pmatrix}.$$

4. Evaluación y Resultados

Según los experimentos llevados a cabo, ListRank-MF logra una mejora en el rendimiento del 15% sobre el método de recomendación colaborativo basado en elementos multimedia y del 10% y 5% sobre el estado del arte constituido por las técnicas CoFiRank-NDGC y CoFiRank-Best, respectivamente. Además, las mejoras son significativas en todas las condiciones de tests evaluadas, aproximadamente. Notar que aunque no hemos especifica-

do el marco de evaluación, este es esencialmente el mismo para todos los algoritmos, por lo que los resultados mencionados resultan relevantes.

En términos aislados, es resaltable que el coeficiente de regularización λ en el ListRank-MF influye en la convergencia de la función de pérdida y controla el sobreajuste. La siguiente figura muestra la relación entre λ y la pérdida, observándose cómo el riesgo de sobreajuste comienza cuando el coeficiente de regularización se encuentra por debajo del valor 0.001.

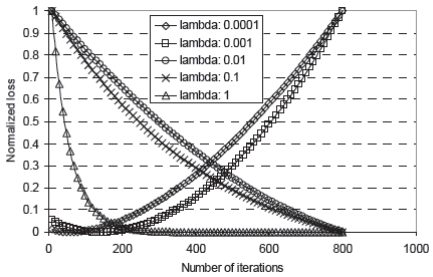


Figura 1. Impacto del coeficiente de regularización sobre la convergencia de la pérdida durante el proceso de aprendizaje.

Finalmente, notar que la optimización de la función de pérdida lleva a la minimización de la pérdida. Como se observa en la siguiente figura, el rendimiento en la clasificación de los elementos multimedia se torna óptimo y convergente cuando se optimiza la función de pérdida.

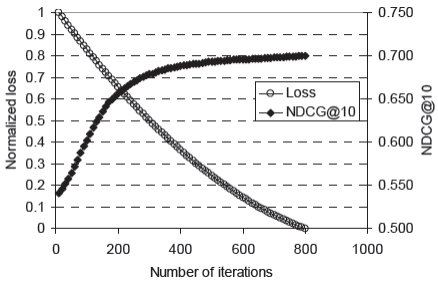


Figura 2. Efectividad del ListRank-MF para lograr la ganancia acumulada descontada normalizada mediante la minimización de la pérdida.

5. Conclusiones

La evaluación llevada a cabo sobre la técnica aquí desarrollada ha demostrado que ListRank-MF mejora la recomendación colaborativa basada en objetos sobre el estado del arte entonces existente. También en la etapa de evaluación fue analizada la complejidad computacional de dicha técnica

verificándose lo expuesto en la introducción, y es que ListRank-MF mantiene una complejidad lineal con la cantidad de ratings observados en la matriz usuario-elemento multimedia dada, pudiendo, por tanto, escalarla con el fin de ser usada en colecciones de contenido realmente grandes propias del mundo real.

6. Referencias

[1] Y. Shi, M. Larson y A. Hanjalic, *List-wise Learning to Rank with Matrix Factorization for Collaborative Filtering.*
[2] R. Salakhutdinov y A. Mnih, *Probabilistic Matrix Factorization.* 2008.