세무사 추천 서비스를 위한 SVD 알고리즘의 RMSE 비교

김원집[1], 허지혜[2], 박세빈[3], 이수민[4], 권은아[5]
한국폴리텍대학 서울강서캠퍼스 빅데이터과
2220110011@office.kopo.ac.kr, floramiss@naver.com, kongsuni3685@gmail.com,
lsms2004@naver.com, chna8137s@gmail.com

RMSE Comparison of SVD Algorithms for Tax Accountant Recommendation Service

Won-Jib Kim[1], Ji-Hye Huh[2], Se-Bean Park[3], Su-Min Lee[4], Eu-Na Kwon[5]
Department of Big Data, Seoul Gangseo Campus of Korea Polytechnics College.

요 약

추천 시스템은 사용자의 선호도를 정확히 파악하는 것이 중요하다. 이를 위해 사용자 데이터를 분석하여 추천을 제공하는 협업 필터링 알고리즘을 활용한다. 하지만 상품의 종류와 고객 수가 많아짐에 따라 사용자 선호도 정확도가 떨어지는 문제점이 있다. 이 문제를 해결하기 위해 제안된 방법은 모델 기반 협업 필터링이며, 이는 고객과 사용자의 정보를 직접적으로 추천하는 대신 모델을 학습시키는데 활용된다. 이에 논문은 추천시스템에서 자주 사용되는 모델 협업 필터링 기반 SVD 모델을 학습 전에 하이퍼 파라미터를 조절하여 모델에 추정 정확도 값인 RMSE를 측정한다.

1. 서론

추천 시스템은 사용자의 데이터를 분석하여 그에 맞는 상품이나 서비스를 추천하는 기술이다. 모든 사용자가 동일한 선호도를 가지고 있는 것이 아니므로, 사용자의 선호도를 정확히 파악하는 것이 중요하다. 이를 위해 사용자 데이터를 분석하여 추천을 제공하는 협업 필터링 알고리즘을 활용한다. 하지만 [1]상품의 종류와 고객 수가 많아짐에 따라 사용자 선호도 정확도가 떨어지는 문제점이 있다. 이 문제를 해결하기 위해 제안된 방법은 모델 기반 협업 필터링이며, 이는 고객과 사용자의 정보를 직접적으로 추천하지 않고 모델을 학습시키는데 활용된다.

이에 논문은 추천시스템에서 자주 사용되는 협업 필터링 기반 SVD 모델을 학습 전에 하이퍼파라미터를 조절하여 모델에 추정 정확도 값인 RMSE를 측정한다. 본 논문의 구성은 다음과 같다. 2장에서는 SVD모델과 RMSE를 알아보고 3장에서는 Python의 Surprise 패키지를 활용한 SVD 모델을 학습 전에 하이퍼파라미터를 조절하여 모델에 RMSE를 측정한다. 마지막 4장에서는 본 연구 결과의 결론을 정리하고 향후 연구 과제를 제시한다.

2. 관련연구

본 장에서는 SVD 모델과 RMSE에 대한 연구를 살펴본다.

2.1. SVD 모델 (Singular Value Decomposition)

SVD 모델은 협업 필터링 기반 모델이다. 사용자, 평점, 아이템 매트릭스를 생성하여 각각 요소들을 특징을 추출한다.

<그림 1> SVD 모델 수식

$$A = U\Sigma V^T$$

<그림 1>은 원본 데이터 행렬 A는 사용자 특성 행렬 U, 대각 행렬 Σ, 그리고 항목 특성 행렬 V^T의 곱으로 분해된다. 대각 행렬 Σ는 특잇값을 포함하며, 이는 데이터의 중요성이나 분산을 나타낸다. 이러한 축소된 데이터를 활용하여 사용자가 평가하지 않은 아이템에 예측값을 계산하고, 순서대로 아이템을 추천한다.

2.2. RMSE (Root Mean Squared Error)

RMSE는 예측 모델의 정확도를 얼마나 잘 예측할 수 있는지 추정하는 값이다.

<그림 2> RMSE 수식

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

실제 사용자의 평점과 시스템의 예측값의 차이를 제곱한 뒤 평균을 구하고 루트를 씌운다. 오차에 대해서 제곱함으로써 1 미만의 오차는 작아지고 그 이상의 오차는 커지는 특징이 있다.

3. 성능 평가

본 장에서는 협업 필터링 기반 SVD 모델의 하이퍼파라미터을 조절하여 RMSE를 비교한다.

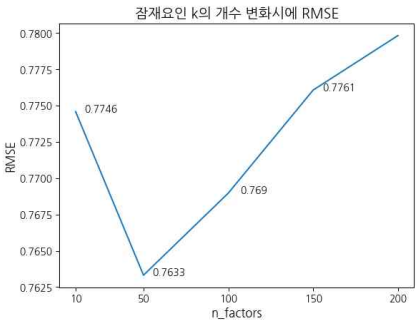
실험은 GroupLens에서 제공하는 MovieLens 데이터의 Small Version으로 600명의 사용자와 9,000개의 영화정보를 활용하였다. [2]학습용 데이터로는 전체 데이터의 75%를 사용하여 Surprise 패키지의 SVD 모델이 학습하고 나머지 데이터 25%를 활용하여 RMSE 측정한다.

< 표 1 > SVD 하이퍼파라미터

파라미터명	비고
n_factors	잠재요인 k의 개수
n_epochs	SGD 수행 시 반복 횟수 (Stochastic Gradient Descent)

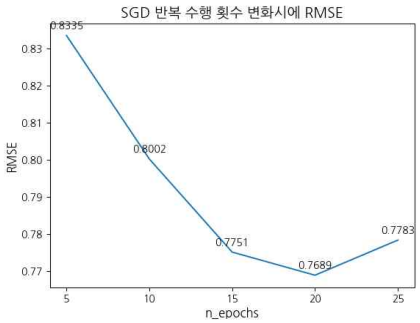
[3]Surprise에서 제공하는 하이퍼파라미터는 <표 1>과 같다.

<그림 3> 잠재 요인 k의 개수 변화 시 RMSE



잠재요인 수를 결정하는 n_factors는 사용자, 아이템, 평점에 영향을 주는 숨겨진 변수를 의미한다. <그림 3>은 잠재요인 k의 수가 50일 때 0.7633으로 최소가 되고 값이 증가할수록 RMSE의 증가를 나타낸다.

<그림 4> SGD 반복 수행 횟수 변화 시 RMSE



SGD 수행 시 반복 횟수를 결정하는 n_epochs는 전체 데이터 세트의 학습 횟수를 결정한다. <그림 4>는 반복횟수가 20일 때 0.7689로 최소가 되고 이후에 RMSE의 증가를 나타낸다.

4. 결론 및 향후 연구과제

본 논문은 추천시스템에서 자주 사용되는 모델 협업 필터링 기반 SVD 모델을 학습 전에 하이퍼파라미터를 조절하여 RMSE를 측정한다. 잠재요인의 k의 수는 50, SGD 수행 반복 횟수의 값은 20 이후에는 학습 데이터의 과적합으로 학습모델의 RMSE가 증가하였다.

향후 모델 성능 개선을 위해서 모델 예측 효율과 정확도를 극대화할 수 있는 모델에 관한 연구가 필요하다.

참고문헌

[1] Jeong Seong-won. "Experimental Study on Product Recommendation Algorithm Using Vector Spatial Model." Graduate School of Soongsil University, 2019. Seoul

[2] https://grouplens.org/datasets/movielens/ MovieLens Latest Datasets

[3] https://surpriselib.com Surprise Package

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.