행렬의 이질성을 고려한 결측치 처리 방법

최기준 황진수 유동현 이우주

인하대학교 통계학과

서론

추천 시스템(Recommender System)에서 사용되는 협업 필터링을 위한 행렬 완성(matrix completion)은 최근 많은 주목을 받았던 문제이다. 그 중, 저차원 행렬에 대한 특이값 분해 방법을 사용하는 SoftImpute (Mazumder et al, 2010)가 행렬 완성 문제에서 우수한 성능을 보여주었다.

추천 시스템은 실제 여러 그룹의 사람으로 구성되어 있으므로, 행렬 내 이질성을 쉽게 관측할 수 있다. 그러나 SoftImpute는 이러한 이질성을 적극적으로 반영하는 방법은 아니다. 따라서 본 연구에서는 행렬 내 이질성이 있을 경우, SoftImpute의 성능을 개선할 수 있는 반복 클러스터링-SoftImpute 방법을 제안해보고자 한다.

연구 동기

협업 필터링

협업 필터링은 사용자들의 선호도에 따라 사용자들의 관심사를 예측하는 방법이다. 근본적으로 사용자들의 과거의 선호도가 미래에도 그대로 유지된다는 가정을 전제로 한다. 이때, 각 상품에 대한 사용자의 선호도를 바탕으로 사용자-상품 행렬을 만들어 아직 접하지 않은 상품에 대한 선호도를 예측한다. 사용자-상품 행렬에 대한 형태는 아래와 같다.

	Item1	Item2	Item3	Item4	Item5	…
User1	2	2			5	…
User2					5	…
User3	1	5	5	4		…
User4		5	4	3	4	…
User5			1	4	4	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

- 사용자가 접하지 않은 부분은 결측값으로 이뤄진 행렬
- 사용자의 선호도가 없는 상품에 대한 부분은 행렬 완성을 통해 예측
- 본 연구에서는 사용자 들 간 이질성이 있는 경우에 대한 처리 방법으로 반복 클러스터링-SoftImpute 방법 제안

기존 분석 방법 소개

SoftImpute – Mazumder et al. (2010)

SoftImpute는 행렬 내 결측이 존재할 때, nuclear-norm 벌칙함수를 고려한 저차원 행렬에 대한 특이값 분해 방법으로 모형의 형태는 아래와 같다.

$$\min \|P_\Omega(X - M)\|_F^2 + \lambda \|M\|_*$$

X: $_{m\times n}$ Matrix
$\|\cdot\|_F$: Frobenius norm
$\|M\|_*$: nuclear norm of M
$P_\Omega(X)$: 행렬 X에서 결측이 아닌 부분

위 모형은 아래의 (1)과 (2), 두 단계를 반복적을 수행하며 최적화한다.

(1) 행렬 X 내 결측인 부분을 현재 추정치 $\widehat{M}$으로 대체:

$$X \leftarrow P_\Omega(X) + P_\Omega^{\perp}(\widehat{M})$$

$P_\Omega^{\perp}(X)$: 행렬 X에서 결측인 부분

(2) 행렬 X에 대한 soft-thresholded SVD를 통해 $\widehat{M}$ 수정:

$$\widehat{X} = UDV^T$$

$$\widehat{M} \leftarrow US_\lambda(D)V^T$$

$S_\lambda(D)$: 대각행렬 D의 대각 성분들인 d_{ii}를 $(d_{ii} - \lambda)_+$로 바꾸는 연산자로, 여기서 $(x)_+ = \max(x, 0)$.

SoftImpute-ALS – Hastie et al. (2014)

$M = AB^T$의 형태를 활용하여 SoftImpute의 수렴 속도를 개선했다.

$$\min_{A,B} \left\|P_\Omega(X - AB^T)\right\|_F^2 + \lambda(\|A\|_F^2 + \|B\|_F^2)$$

X: $_{m\times n}$ Matrix
$A_{m\times r}$, $B_{n\times r}$: Rank $r \le \min(m, r)$

제안 방법

반복 클러스터링-SoftImpute

사용자 간 이질성을 고려한 행렬 완성 방법으로 클러스터링과 SoftImpute의 반복을 제안한다. 자세한 방법은 아래 (1)-(4)의 순서와 같이 진행된다.

(1) 초기 시행 시 행렬 $X_{m\times n}$의 결측은 0으로 채운 후, 행 기준으로 계층 클러스터링

(2) 최적의 클러스터 개수 K 결정:

$$K = \text{argmax}_i(s_i) = \max_i \frac{b_i - a_i}{\max(a_i, b_i)}$$

a_i: i번째 행과 같은 군집 내 다른 관측치와의 평균거리
b_i: i번째 행과 다른 군집 간 최소거리

(3) 각 클러스터 별로 SoftImpute-ALS:

$$\min_{A,B} \left\|P_\Omega(X_{(i)} - A_{(i)}B_{(i)}^T)\right\|_F^2 + \lambda(\|A_{(i)}\|_F^2 + \|B_{(i)}\|_F^2)$$

$X_{(i)}$: i번째 클러스터에 포함되는 행들로 구성된 행렬 $(i = 1, \dots, K)$

(4) 500번 혹은 행렬의 변화가 극히 적을 때까지 (1)-(3) 과정 반복

Simulation Study

제안된 방법의 성능 향상을 확인하기 위해 다양한 이질적인 상황 하에서 SoftImpute와 제안된 방법의 결측치 예측력에 대한 수치 연구를 실시하였고, 기존의 방법과 제안한 방법을 적용하여 예측값의 RMSE를 비교하였다.

Scenario 1) Homogeneity (n = 400, p = 50, group = 1)

Simulation Data: $X_{400\times 50} = U_{400\times 2}\,\Lambda_{2\times 2}\,V_{2\times 50}^T$

	Missing Proportion					
Method	20%	40%	60%	80%	90%	95%
SoftImpute	0.217	0.699	2.275	2.174	9.95	29.708
클러스터링 - SoftImpute	0.283	0.723	2.198	2.151	7.226	28.376

Scenario 2) Heterogeneity (n = 400, p = 50, group = 4)

Simulation Data: $X_{400\times 50} = (X_1, X_2, X_3, X_4)$,

$X_i = U_{100\times 2}\,\Lambda_{2\times 2}\,V_{2\times 50}^T, (i = 1, 2, 3, 4)$

	Missing Proportion					
Method	10%	30%	50%	80%	90%	95%
SoftImpute	0.778	2.028	2.84	12.213	198.081	1180.311
클러스터링 - SoftImpute	0.112	0.331	1.067	5.095	131.392	602.076

Scenario 3) Heterogeneity (n = 2000, p = 50, group = 4)

Simulation Data: $X_{2000\times 50} = (X_1, X_2, X_3, X_4)$,

$X_i = U_{500\times 2}\,\Lambda_{2\times 2}\,V_{2\times 50}^T, (i = 1, 2, 3, 4)$

	Missing Proportion					
Method	10%	30%	50%	80%	90%	95%
SoftImpute	0.852	1.601	31.997	75.552	140.602	139.047
클러스터링 - SoftImpute	0.119	0.44	7.761	55.116	115.798	71.695

실제 데이터) Movielen 100K data

	Missing Proportion				
	84%	87%	86%	84%	94%
Method	Size (300, 200)	(800, 600)	(900, 500)	(900, 1200)	(943, 1682)
SoftImpute	0.331	0.281	0.285	0.336	0.320
클러스터링 - SoftImpute	0.285	0.280	0.272	0.321	0.285

Conclusion

본 연구에서는 사용자 간 이질성을 고려한 방법으로 반복 클러스터링-SoftImpute 방법을 제안하였다. 모의실험 결과를 통해서 행렬 내 이질성이 존재할 경우에는 기존 방법보다 우수한 성능을 보이는 것을 확인하였다. 동시에 이질성이 없는 경우에도 제안한 방법이 기존 방법에 비해 큰 차이가 없었다. 추가적으로 실제 데이터 Movielen을 통해서 제안한 방법의 우수함을 알 수 있었다.

Reference

[1] R Mazumder, T Hastie, R Tibshirani, Spectral Regularization Algorithms for Learning Large Incomplete Matrices, Journal of Machine Learning Research, 11:2287–2322, 2010

[2] T Hastie, R Mazumder, J Lee, R Zadeh, Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares, Journal of Machine Learning Research, 16:3367-3402, 2015