Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 441~450

Improvement of Collaborative Filtering Algorithm Using Imputation Methods[1)]

Hyeong Chul Jeong[2)] · Minjung Kwak[3)] · Hyunju Noh[4)]

Abstract

Collaborative filtering is one of the most widely used methodologies for recommendation system. Collaborative filtering is based on a data matrix of each customer's preferences and frequently, there exits missing data problem. We introduced two imputation approach (multiple imputation via Markov Chain Monte Carlo method and multiple imputation via bootstrap method) to improve the prediction performance of collaborative filtering and evaluated the performance using EachMovie data.

Key words : 협업필터링, 결측값, 다중대치, 붓스트랩, MCMC, EM 알고리즘

1. 서 론

E-CRM을 활용하는 기업의 서비스 형태 중 하나는 사용자가 웹 사이트에 접속했을 때 사용자에게 적절한 항목(item)을 추천하는 것이다. 이러한 추천 시스템은 고객으로 하여금 수많은 정보 중에 효율적으로 자신에게 필요한 정보를 빠른 시간에 찾을 수 있게 함으로 고객의 만족도를 높여 기업이익을 극대화 시킬 수 있는 효과가 있다.

추천시스템에 사용되는 방법으로 크게 내용기반 필터링(content-based filtering)과 협업 필터링(collaborative filtering)으로 구분할 수 있다. 특히, 협업필터링 방법은 내용기반 필터링 방법의 단점을 보안한 방법으로 동일한 정보를 필요로 하거나 동일한 성향을 지닌 사람들을 연결하여 같은 영역(domain) 내에 있는 항목에 대해 평가 결과를 공유함으로써 사용자들이 항목에 대해 좀더 나은 결정을 하도록 하는 것이다

1) 본 연구는 2001년 평택대학교 E-비지니스 연구소 지원에 의하여 수행되었음.
2) 경기도 평택시 용이동 111 평택대학교 정보통계학과 조교수
E-mail: jhc@ptuniv.ac.kr
3) 경기도 평택시 용이동 111 평택대학교 정보통계학과 조교수
4) 서울시 동대문구 청량리동 207-43 한국과학기술원 경영대학원

(David et al, 1992; Joseph et al, 1999; Paul et al, 1994). 그런데, 사용자로부터 직접 받은 선호도 정보를 바탕으로 하기 때문에, 사용자들의 응답률이 낮거나, 데이터가 희소한 경우 사용자의 선호도를 부정확하게 반영하여 추천 시스템의 정확도가 저하될 수 있다. 즉 협업필터링의 가장 큰 문제점 중에 하나는 자료에 결측치가 많이 발생한 경우라 할 수 있다. 결측치가 많이 발생한 경우, 이러한 결측치를 대치하여 협업필터링을 실시하는 것이 추천시스템의 정확도 향상에 보다 효율적이라 하겠다. 여기서 결측값을 대치하는 몇 가지 편의적인 방법으로 관측된 값에서 표본추출을 하여 결측값을 대치하는 방법, 평균값으로 결측값을 대치하는 방법, 회귀분석에 의한 대치방법 등이 있다. 이와같은 대치방법에 대해 개선된 방법으로 Little and Rubin(1987)은 EM 알고리즘(Dempster et al, 1977)을 이용한 대치방법을 권장하고 있으며, 이는 결측치에 대한 고전적 방법으로 알려져 있다.

그러나, 앞에서 언급한 대치방법들은 결측값에 대해 하나의 값을 대치하는 단일대치(single imputation) 방법으로, 단일대치는 추정치나 추정치의 분산을 왜곡시킬 수 있다(Rubin, 1996). 예를 들어 사용자간의 상관관계를 추정하는 경우, 예측한 값이 실제의 상관관계 값보다 더 부풀려지게 추정되는 경향이 존재한다. 또한 단일대치에 의해 얻어진 자료로부터의 추정은 결측값의 비율이 높을수록 정확도에 문제가 발생 할 수 있으며, 단일대치에 의한 측정값의 불확실성에 대한 척도(measure of uncertainty)를 측정할 수 없다. 그러므로 본 연구에서는 협업필터링 과정에서 단일대치의 여러 문제를 보완한 다중대치방법으로 Markov Chain Monte carlo(MCMC) 방법을 활용한 다중대치와 붓스트랩 다중대치를 소개하고자 한다.

본 연구의 2절에서는 기존의 협업필터링 알고리즘에 대해 간단히 소개하고, 3절에서는 결측값이 존재할 때 협업필터링 알고리즘을 개선할 수 있는 MCMC를 활용한 베이지안 관점의 다중대치방법과 빈도주의적 관점의 붓스트랩 다중대치방법을 소개하고자 한다. 4절에서는 실제자료를 통해 두 방법을 비교하였다.

2. 협업필터링

협업필터링은 사용자들의 선호도에 대한 자료를 기반으로 새로운 사용자가 관심을 가질 것으로 생각되는 항목(상품, 광고, 웹 페이지 등)을 추천해 주는 기법이다. 규칙기반 필터링(rule-based filtering)이나 매칭 에이전트(matching agent) 등이 항목자체의 속성 정보를 사용하여 사용자에게 추천하는 것과 달리 협업필터링은 항목에 대한 다른 사용자들의 선호도 점수를 기반으로 하기 때문에 협업이라는 용어를 사용하게 된다. 이러한 협업필터링은 Jonathan et al(1999), Joseph et al(1997) 등에 의해 추천 시스템으로 유용하게 활용되었다.

협업필터링은 일반적으로 가중치부여(weighting), 경계값 선정(thresholding), 예측(predicting) 의 3단계로 구성된다.

2.1 가중치부여

고객간의 유사도 가중치를 구하기 위해서 사용되는 유사도 기준값으로는 대표적으로 상관계수나 벡터유사도(vector similarity) 등이 사용된다. 본 연구에서는 Pearson

상관계수를 사용하여 유사도 가중치를 계산하였다. Pearson 상관계수를 사용할 경우 사용자 a와 사용자 i의 유사도 가중치는 다음과 같다.

$$w(a,i)=\frac{\sum\ (v_{a,j}-\overline{v_a})(v_{i,j}-\overline{v_i})}{\sqrt{\sum_j\ (v_{a,j}-\overline{v_a})^2\sum_j\ (v_{i,j}-\overline{v_i})^2}} \tag{2.1}$$

여기서 $v_{a,j}$는 사용자 a가 항목 j에 대해서 보여준 선호도이고 $\overline{v_a}$는 사용자 a가 선호도를 입력한 항목들에 대한 선호도 평균값이다. j는 사용자 a와 i가 공통으로 선호도를 입력한 항목들에 해당된다.

2.2 경계값 선정

사용자간의 유사도 가중치의 값을 계산한 후에는 특정 사용자의 특정 항목에 대한 선호도를 예측하기 위해 몇 명의 이웃(neighborhood)을 사용할 것인가를 결정해야 한다. 예측에 사용될 이웃들을 결정하기 위해서 Thresholding과 Best-n-neighborhood 방법을 사용할 수 있다. Thresholding은 사용자간의 유사도 가중치가 어느 정도의 값 이상인 이웃들만을 사용해서 예측하도록 제안하는 방법이고 Best-n-neighborhood는 특정 사용자와 유사한 n명의 이웃을 사용해서 예측하도록 제안하는 방법이다. 또한 위의 두 방법을 조합하여 유사도 가중치가 어느 정도 값 이상인 이웃들 중 n명을 선택하는 방법을 사용할 수 있다.

2.3 예측

가중치와 이웃이 선정되면 추천을 위한 예측통계량이 결정되어야 한다. 예측통계량은 다양하게 존재하나 본 연구에서는 이웃들의 선호도와 선호도 평균과의 거리를 이웃들과의 유사도로 가중평균 함으로써 특정 사용자의 특정 항목에 대한 선호도를 예측하는 평균으로부터 편차(deviation-from-mean)방법을 사용하여 예측하였다. 이는 다음 식처럼 표현된다.

$$p_{a,j}=\overline{v_a}+\frac{\sum_{i=1}^{n} w(a,i)(v_{i,j}-\overline{v_i})}{\sum_{i=1}^{n} |\,w(a,i)\,|} \tag{2.2}$$

여기서, $p_{a,j}$는 사용자 a의 항목 j에 대한 선호도를 예측한 값이고, $\overline{v_a}$는 사용자 a의 선호도 평균값이다. $w(a,i)$는 사용자 a와 사용자 i의 유사도 가중치이고 n은 사용자 a와 다른 사용자들 간의 유사도가 0이 아닌 사용자수이다. 유사도 가중치 $w(a,i)$는 앞에서 설명한 것과 같은 Pearson 상관계수로 구하게 된다.

3. 대치를 이용한 협업필터링

3.1 MCMC 방법을 활용한 다중대치

본 절에서는 선호도 조사에서 발생된 결측치를 MCMC 알고리즘을 통하여 다중대치를 실시한 후 협업필터링을 사용하는 방법을 소개하고자 한다. 다중대치를 실시하기 위해서는 (1) 결측값의 모형은 결측값이 아닌 관찰값에 의존한다는 랜덤결측(missing at random : MAR) 가정 (2) 자료의 분포모형 (3) 모수에 대한 사전 분포 등의 가정이 요구된다.

관찰자료를 Y_{obs}, 결측자료를 Y_{mis}, 대치된 자료를 Y_{imp}, 반응에 대한 지시값을 R이라 놓으면, MAR 가정에 의해 R의 분포는 관찰자료에만 종속되는 분포를 따르게 된다. 즉 R의 분포는 다음 식으로 표현될 수 있다.

$$P(R|\ Y_{obs},\ Y_{mis}) = P(R|\ Y_{obs})$$

자료 $Y=(Y_{obs}, Y_{mis})$에 결측값을 대치한 완전한 자료는 $P(Y_{mis}|\ Y_{obs})$의 분포로부터 생성할 수 있다. 또한 관찰값에 기초한 결측값의 분포는 다음과 같다.

$$P(Y_{mis}|\ Y_{obs}) = \int P(Y_{mis}|\ Y_{obs}, \theta)P(\theta|\ Y_{obs})\,d\theta \qquad (3.1)$$

여기서 $P(\theta|\ Y_{obs})$는 관찰값 Y_{obs}에 의한 모수 θ의 사후분포이다. 모수 θ에 대한 사전분포 $\pi(\theta)$와 우도함수 $L(\theta|\ Y_{obs})$가 주어지면 $P(\theta|\ Y_{obs}) \propto L(\theta|\ Y_{obs})\pi(\theta)$이다. 그리고, 위의 $P(Y_{mis}|\ Y_{obs})$를 유도하기 위해서는 자료의 분포에 대한 가정과 모수 θ에 대한 사전분포가 요구됨을 알 수 있다. 식 (3.1)에 기초하여 결측값 대치를 하기 위해서는 MCMC 방법이 사용될 수 있다. MCMC 알고리즘을 통한 자료 생성과정은 다음과 같다.

[단계1] $Y_{imp}^{(t+1)} \sim P(Y_{mis}|\ Y_{obs}, \theta^{(t)})$ 에서 Y_{imp}를 랜덤추출한다.

[단계2] $\theta^{(t+1)} \sim P(\theta|\ Y_{obs}, Y_{mis}^{(t+1)})$ 에서 모수 θ를 랜덤추출한다.

[단계3] [단계1]과 [단계2]를 반복하여 다음의 마코프 연쇄를 생성한다.

$$Y_{imp}^{(1)}, \theta^{(1)},\ Y_{imp}^{(2)}, \theta^{(2)}, \cdots,\ Y_{imp}^{(t)}, \theta^{(t)}, \cdots,$$

여기서, [단계3]의 계열은 $P(Y_{mis}, \theta|\ Y_{obs})$에 분포수렴(in distribution)하는 것으로 알려져 있다.

이러한 MCMC 방법을 통한 자료증폭(data augmentation)의 과정으로 다양한 크기의 대치자료를 생성할 수 있게 된다. 이제, MCMC 방법을 통해 얻어진 마코프 연쇄 자료계열에서 길이가 t 계열인 $Y_{imp}^{(t)}, Y_{imp}^{(2t)}, \cdots, Y_{imp}^{(mt)}$ 를 선택함으로 완전히 대치된 m개의 자료를 생성할 수 있다. 만일, m이 1이라면 1개의 결측치 자료를 유도한 것이라 할 수 있다. 그리고 m개의 완전한 자료집합으로부터 협업필터링을 각각 실시한다. 즉, 사용자 a의 항목 j에 대한 선호도를 예측하고자 하면, 식(2.2)를 통하여 m개의

예측값 $p_{a,j}^{(1)}, p_{a.,j}^{(2)}, \cdots, p_{a,j}^{(m)}$를 얻을 수 있으며, 이로부터 선호도 $p_{a,j}$의 예측 결과는 $\hat{p}_{a,j} = \sum_{i=1}^{m} p_{a,j}^{(i)}/m$ 로 주어진다.

본 연구에서 데이터의 분포모형으로 다변량 정규분포를 사용하였다. 다변량 정규분포를 가정하는 경우 실제 선호도 자료는 순서형이므로 각 구간을 설정하여 구간에 포함되는 값을 지정할 수 있다. 또한 다변량 정규분포모형은 순서형 자료와 같은 범주형자료에도 비교적 잘 적합되는 것으로 나타난다. 그리고 모수에 대한 사전분포로는 무정보(noninformative) 분포를 가정하였다. 실질적으로 결측값대치에 있어서 $\pi(\theta)$의 선택에 민감하지 않으며 일반적으로 무정보 분포를 사용하는 것이 보통의 모형에 잘 적합된다고 알려져 있다(Schafer, 1997).

3.2 붓스트랩 방법을 활용한 다중대치

Efron(1994)은 결측자료에 대해 빈도주의적 접근방법인 붓스트랩방법을 활용하였다. 즉 EM 알고리즘을 사용하여 단일대치 된 자료에 대해 붓스트랩 방법을 적용함으로 다중대치의 효과를 유도하고자 하였다.

결측자료에 대해서 붓스트랩 방법의 적용은 크게 두 가지로 구분할 수 있다. 원자료 행렬Y가 결측값을 포함할 때 우선 결측값을 포함한 상태로 붓스트랩을 하고 결측값을 포함한 붓스트랩 자료행렬에 대해 각각 결측값을 채워 넣는 방법(Bootstrap and Imputation)과 원자료에 결측값을 채워 넣은 후 붓스트랩을 하는 방법(Imputation and Bootstrap)을 생각할 수 있다.

결측치가 포함된 자료 $Y = (Y_{obs}, Y_{mis})$에서 대치 후 붓스트랩을 하여 협업필터링을 하는 방법은 다음과 같다.

[단계1] $Y = (Y_{obs}, Y_{mis})$에 EM 방법으로 결측치를 단일대치한 Y_{imp} 자료를 얻는다.

[단계2] Y_{imp} 로부터 B 개의 붓스트랩 자료집합 $Y_{imp}^{*(1)}, Y_{imp}^{*(2)}, \cdots, Y_{imp}^{*(B)}$ 를 발생한다.

[단계3] 각각의 붓스트랩 자료로부터 협업필터링을 실시하여 $p_{a,j}^{(1)}, p_{a.,j}^{(2)}, \cdots, p_{a,j}^{(B)}$ 로부터 선호도를 $\sum p_{a,j}^{(i)}/B$ 로 예측한다.

붓스트랩 후 대치를 하여 협업필터링을 실시하는 방법은 다음과 같다.

[단계1] $Y = (Y_{obs}, Y_{mis})$에 대해 주어진 결측치를 포함한 상태로 붓스트랩 자료집합 $Y_{mis}^{*(1)}, Y_{mis}^{*(2)}, \cdots, Y_{mis}^{*(B)}$을 발생한다.

[단계2] $Y_{mis}^{*(i)}, i = 1, \cdots B$ 의 붓스트랩 자료 각각에 대해 EM 방법을 통해 결측치를 단일대치한 $Y_{imp}^{*(i)}, i = 1, \cdots, B$ 를 얻는다.

[단계3] 각각의 붓스트랩 자료 $Y_{imp}^{*(i)}$ 로부터 협업필터링을 실시하여 $p_{a,j}^{(1)}, p_{a.,j}^{(2)}, \cdots, p_{a,j}^{(B)}$ 로부터 선호도를 $\sum p_{a,j}^{(i)}/B$ 로 예측한다.

위의 붓스트랩 방법에서 대치 후 붓스트랩 하는 방법은 처음 대치 결과에 크게 의존하는 단점이 있으므로 대치 후 붓스트랩보다는 붓스트랩 후 대치 방법이 다소 효율적이라고 할 수 있다.

4. 자료분석

4.1 자료구조

본 절에서는 실제자료를 통하여 앞에서 언급한 대치방법들의 효율성을 비교하고자 한다. 본 연구의 실험을 위해서 사용한 자료는 Compaq 연구소에서 18개월 동안 협업필터링 알고리즘을 연구하기 위해 수집한 EachMovie 자료이다(McJones;1997). Eachmovie 자료에는 총 72,916명의 사용자와 1,648종류의 영화가 존재하며 영화에 대해 사용자가 보인 2,811,983번의 선호도 정보가 존재한다. 선호도는 1(가장 싫은 영화)에서 6(가장 좋은 영화)인 6단계로 표시된다. 72,916명의 사용자중 한 번이라도 선호도를 보인 사용자는 61,263명이고 1,648종류의 영화 중 한 번이라도 선호도가 표시된 영화는 1,623개이다. 선호도가 입력된 모든 자료의 히스토그램은 (그림4.1)의 (A)에 주어져있다.

분석을 위하여 61,263명의 사용자에서 임의로 31명의 사용자를 선택한 후 25명은 training에 활용하고 나머지 6명은 testing에 활용하였다. 25명 training 사용자 전체를 분석에 활용하고자 (2.2) 절의 이웃을 선택하여 협업필터링을 실시하는 방법을 사용하지 않았다. 또한 1,628종류의 영화를 결측치 수준이 25%, 50% 그리고 75%인 세 그룹으로 분류하였다. 여기서 선정된 사용자들의 점수 분포가 전체경향을 나타내는 그림(A)처럼 영화에 대해 보통과 보통이상의 긍정적 반응을 보이는 약간의 정규분포 형태를 따르리라 기대할 수 있다. 그런데, 대치방법이 분포에 영향을 받지 않을 조건을 만들기 위해 40%의 training 사용자는 그림(B)와 그림(C)의 형태처럼 일반적인 경향과 차이가 있는 사용자를 선택하였다. 또한 Testing 집단으로 선정된 6명도 각각 그림(D), (E) 그리고 (F)의 경향을 따르는 사용자를 선택하였다. 이와같은 경향은 정규성에서 크게 벗어나 있다고 하겠다.

4.2 평가결과

본 자료분석에서는 기존의 협업필터링 방법과 앞에서 언급한 대치방법을 사용한 협업필터링의 효율성을 3 수준의 결측치별로 비교하였다. 또한 기존의 협업필터링(CF) 방법과 계열 $t = 50$에서 다중대치를 각각 1회(EM), 3회(MI3), 5회(MI5), 7회(MI7), 10회(MI10) 하여 협업필터링을 실시하는 방법과 EM 알고리즘을 사용하여 단일대치된

자료로부터 100회의 붓스트랩 자료에 의한 협업필터링 방법(EMBT) 그리고 결측치가 포함된 자료를 100회 붓스트랩 한 후 EM알고리즘으로 단일대치한 후 협업필터링을 실시하는 방법(BTEM) 등 총 8가지 방법의 효율성을 비교하였다. 평가를 위해서는 평균절대오차(mean absolute error)와 평균절대오차의 분산 그리고 선호예측도(positive prediction)를 측정하였으며, S-plus로 프로그램 하였다.

(그림 4.1) EachMovie 자료 전체와 Training 집단 그리고 Testing 집단에 대한 히스토그램

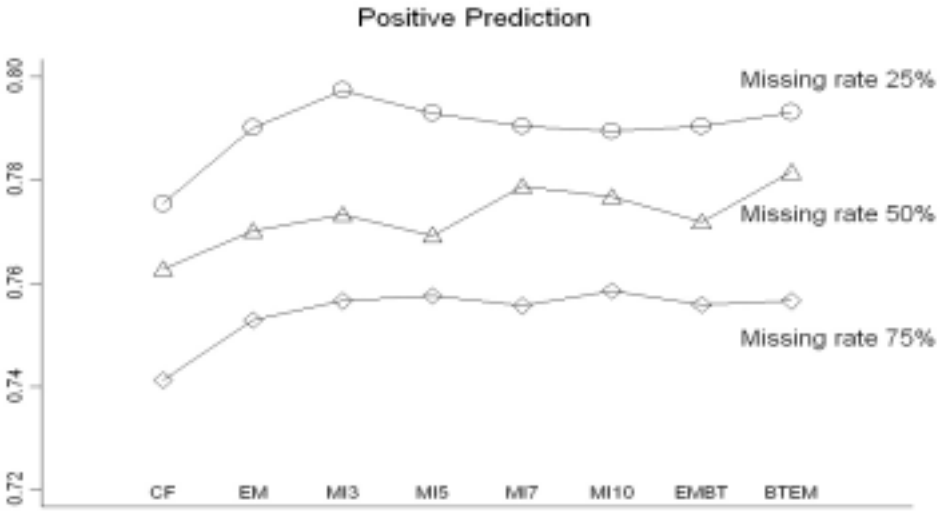
결측률	CF	EM	MI3	MI5	MI7	MI10	EMBT	BTEM
25%	0.87768	0.84487	0.84400	0.84521	0.84422	0.84457	0.85042	0.83331
	(0.39082)	(0.40109)	(0.39026)	(0.38592)	(0.38537)	(0.38552)	(0.38883)	(0.38489)
50%	0.98725	0.94800	0.93184	0.92390	0.92867	0.92648	0.94891	0.92402
	(0.52838)	(0.53734)	(0.50896)	(0.50922)	(0.50910)	(0.50667)	(0.56031)	(0.49101)
75%	1.19344	1.09628	1.07459	1.07139	1.06629	1.06125	1.09781	1.06215
	(0.70123)	(0.66125)	(0.62417)	(0.62279)	(0.61762)	(0.62159)	(0.66080)	(0.62213)

(표 4.1) 결측값의 비율에 따른 각 방법들의 평균절대오차와 평균절대오차의 분산

(표 4.1)은 세 수준의 결측값의 비율에 따라 8가지 방법들의 협업필터링에 의한 예측방법의 평균절대오차와 평균절대오차의 분산을 비교한 표이다. (표 4.1)로부터 다중대치를 사용한 예측방법이 기존의 협업필터링을 사용하는 방법보다 모두 평균절대오차가 작게 나타나므로 다중대치를 이용한 방법의 예측력이 우수함을 알 수 있다. 또한 MCMC를 활용한 다중대치보다는 붓스트랩 후 대치를 사용하는 BTEM 방법이 대체적으로 평균절대오차가 작음을 볼 수 있다.

EachMovie 자료에서 결측값이 25%인 경우는 응답자가 해당 영화에 대한 충성도나 호응도 및 관심도가 높은 영화라 할 수 있다. 이와같은 경우에 단순히 협업필터링을 사용하여 선호도를 예측하는 것보다는 다중대치를 사용하는 것이 선호도 예측에 대한 정확도가 높으며, 붓스트랩 후 대치(BTEM) 방법의 선호도 예측이 가장 뛰어난 수준임을 볼 수 있다. 이러한 경향은 결측수준 50%와 75%에서도 비슷하게 나타나며, 결측수준이 25%에서 50% 그리고 75%로 높아질수록 효율성이 CF에 대비하여 BTEM을 사용하였을 때, 각각 5.1%, 6.4%, 11.0% 증가함을 볼 수 있다. 또한, 결측수준 25%와 50%에서 평균절대오차의 분산도 CF보다 다중대치방법이 그리고 다중대치보다는 BTEM 방법 순으로 작아짐을 볼 수 있겠다. 다중대치 횟수가 1회인 단일대치(EM)와 3회, 5회, 7회 그리고 10회인 경우를 살펴보면, 결측수준 25%에서 MCMC를 이용한 다중대치 3회의 평균절대오차가 작으며, 결측수준이 50%인 경우에는 다중대치 5회, 그리고 결측수준이 75%인 경우에는 다중대치를 10회 실시할 때가 평균절대오차가 낮음을 볼 수 있다. 즉 결측수준이 높아질수록 다중대치 횟수가 더 많이 필요하나 일반적으로 5회에서 7회 정도면 충분하리라 여겨진다. 단일대치 후 붓스트랩을 실시하는 EMBT 방법은 EM에 의한 단일대치의 자료구조에 크게 의존하므로 평균절대오차가 EM의 경우 보다 다소 높으며, 평균절대오차의 분산 역시 다른 방법들에 비해 다소 높은 수준에서 형성되는 단점이 있음을 볼 수 있다. 결론적으로 BTEM 방법은 MCMC를 사용한 MI 방법 중 최적의 대치회수 경우와 비슷한 경향을 보이고 있으며 EMBT 방법은 대체적으로 EM 보다 효율성이 떨어지는 방법임을 볼 수 있다.

(그림 4.2)는 선호도를 2단계로 표시하여 선호도가 3이하면 비선호로 4이상이면 선호로 나누었을 때 선호예측도를 보여준다. 여기서 선호예측도란 알고리즘에 의해서 추천된 항목 중 사용자가 실제로 선호한 항목의 비율을 의미한다. (그림 4.2)에서 결측값의 비율이 낮을수록 선호예측도가 높음을 볼 수 있다. 또한 결측값이 낮은 수준인 25%에서 CF 방법의 선호예측도가 다른 방법에 비해 가장 낮으며, 결측값 비율이 높아질수록 기존 CF 방법의 선호예측도가 더 아래로 쳐지는 형태를 보임을 볼 수 있다. 이는 결측값의 비율이 높으면 기존의 CF방법은 대체로 선호도를 평균값으로 예측하기 때문이라고 짐작 할 수 있겠다. 즉 결측값이 높은 수준에서 CF 방법에 의한 선호도는 대체로 3 또는 4로 예측되었다. 더구나 선호하지 않는 영화는 선호하는 영화에 비해 결측값이 많이 나타날 가능성이 높으므로 결과를 왜곡하고 예측값이 높게 추정될 가능성이 존재한다고 하겠다. 각각의 결측치 수준에서 보면 25% 수준에서는 MI3, 50% 수준에서는 MI7과 BTEM, 그리고 75% 수준에서는 MI10이 우수한 방법이라 하겠다. 선호예측도와 비슷한 개념인 민감도와 특이도에서도 비슷한 경향이 나타나므로 여기서는 생략하기로 한다.



(그림 4.2) 각 결측수준에 따른 선호예측도

5. 결론

본 연구는 전자상거래에 있어 개인화 된 추천시스템을 구축하는데 주로 사용되고 있는 협업필터링(collaborative filtering)의 알고리즘을 보완하고 실제 사례에서 적용을 다루었다. 개인화 서비스를 위한 알고리즘 중 하나인 내용기반 필터링의 경우, 개인 사용자들의 경험에 주로 의존하기 때문에 추천하는 상품에 한계를 가지는 단점을 가지고 있는 반면, 다른 사용자들의 선호도를 기반으로 하는 협업필터링은 제품 선택에 있어서 내용기반 필터링의 한계를 넘어서는 장점을 가지고 있다. 그러나, 사용자들이 선호도를 입력하지 않는 한, 분석의 기반이 되는 사용자-선호도 행렬에 결측값이 발생하여 선호도가 유사한 사용자를 선택하거나 추천할 수 있는 항목에 대한 제약이 발생하는 등 분석 자료의 희소성 문제로 추천시스템 모형의 예측능력이 저하되는 문제가 발생한다. 따라서 이러한 희소성 문제는 협업필터링 적용과정에서 해결해야 할 주요 선행 과제 중 하나라고 할 수 있겠다. 본 연구에서는 결측값을 추정할 수 있는 방법으로 다중대치 방법과 붓스트랩 방법을 적용하여, 자료의 희소성 문제를 해결하고자 하였다. 단순히 결측값을 제거하거나, 가중 평균 등을 이용하는 것보다 EM 및 MCMC 방법 등을 사용하여 결측값을 보완한 완전한 자료를 기반으로 협업필터링 실시하여, 추천시스템의 예측능력을 보다 개선할 수 있을 것으로 기대한다. 자료분석 결과를 종합하면, 기존의 협업필터링과 비교하여 다중대치를 이용한 방법과 붓스트랩대치 방법의 예측력이 높은 것으로 나타났다. 그런데, 실제 효율적인 측면에서는 붓스트랩 방법은 다중대치에 비해 계산시간이 많이 걸린다는 단점이 있다. 다중대치 역시 매우 많은 횟수의 대치가 요구된다면 붓스트랩 방법과 비교하여 특별한 잇점이 없을 수 있다. 그런데 다중대치는 결측치 비율이 아주 높은 경우에도 7회 혹은 10회 정도면 충분한 것으로 나타난다(Schafer, 1997). 이는 다중대치법이 지니는 계산적 장점이라고 할 수 있겠다. 이와같은 잇점에 따라 다중대치방법이 비용과 실시간 처리 측면

에서 붓스트랩 대치방법보다 실제적으로 선호 될 수 있겠다.

본 연구에서는 제안된 알고리즘을 포함한 다양한 협업필터링 알고리즘의 예측능력을 결측값의 수준에 따라 자료분석 하였으나, 추후 실제 전자상거래 사이트에 적용시 사이트 성격에 따라 발생하는 데이터의 특성에 따라 보다 다양한 방법론들이 제안되리라 기대한다.

참고문헌

1. David, G., David, N., Brian, M. O. and Douglas, T. (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM,* 35(12), 61-70.
2. Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, series B*, vol. 39, 1-38.
3. Efron, B (1994). Missing data, Imputation, and the Bootstrap, *Journal of the American Statistical Association,* 89, 463-479.
4. Jonathan, L. H., Joseph, A. K., Al, B. and John, R. (1999). An Algorithmic Framework for Performing Collaborative Filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval,* 230-237, Berkeley, CA.
5. Joseph, A. K., Bradley, N. M., David, M., Jonathan, L. H., Lee, R. G. and John, R. (1997). GroupLens : Applying Collaborative Filtering to Usenet News. *Communications of th ACM,* 40(3), 77-87.
6. Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data,* Wiley & Sons.
7. McJones, P. (1997). Eachmovie Collaborative Filtering Data set. http://www.research.digital.com/SRC/Eachmovie, DEC Systems Research Center.
8. Paul, R., Neophytos, I., Mitesh, S., Peter, B. & John, R. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM CSCW'94 Conference on Computer Supported Cooperative Work,* 175-186.
9. Rubin, D. (1996). Multiple Imputation After 18+ Years, *Journal of the American Statistical Association, Vol. 91.* 473-489.
10. Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall.
11. Schafer, J. B., Joseph, K. & John, R. (1999). Recommender Systems in E-Commerce. In *Proceedings of the ACM Conference on Electronic Commerce.*

[2003년 3월 접수, 2003년 6월 채택]