creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사 학위논문

제한볼츠만머신을 이용한 추천시스템 소개 및 자료분석

Introduction of Restricted Boltzmann Machines for recommender system and real data analysis

2018년 2월

서울대학교 대학원

통계학과

정 바 램

국문초록

초기의 추천시스템은 전반적인 매출의 실태를 이용하여 사용자에게 상품을 추천하였다. 전체평균에 의존한 추천은 개인의 개성과 상품의 특성을 반영하지 못하기 때문에 매출상승을 위한 도구로 효율이 떨어지기 마련이다. 이러한 문제를 해결하기 위해서, 등장한 협업필터링(CF:Collaborative Filtering) 방법은 사용자 기반 방법(User based CF)과 상품 기반 방법(Item based CF)으로 사용자에게 새로운 상품을 추천한다. 그러나 사용자-상품 행렬의 희소성(sparse) 문제로 정확한 상품 추천이 어렵다는 한계점을 가진다. 이때, 기계학습(Machine Learning)을 위해 사용되는 모형으로, 비지도학습(Unsupervised Learning)의 한 종류인 제한볼츠만머신(Restricted Boltzmann Machines)을 적용하여 상품 추천 능력을 향상 시킬 수 있다. 본 논문에서는 제한볼츠만머신을 이용한 협업필터링방법 알고리즘을 소개하고 실제 한국 영화 평점 자료를 이용해 분석 및 적용해본다.

주요어 : 추천시스템, 협업필터링, 사용자 기반 방법, 상품 기반 방법, 제한볼츠만머신, 기계학습, 비지도학습

학　번 : 2016-20275

Contents

국문초록 **ii**

1 서론 **1**

2 제한볼츠만머신 **3**

2.1 제한볼츠만머신의 구조 3

2.2 제한볼츠만머신의 확률분포 5

2.3 제한볼츠만머신의 모수 및 학습 알고리즘 6

3 추천시스템을 위한 제한볼츠만머신 **9**

3.1 사용자 기반 제한볼츠만머신 10

3.2 상품 기반 제한볼츠만머신 13

3.3 하이브리드 제한볼츠만머신 15

4 시뮬레이션 **17**

4.1 데이터 설명 17

5 결론 및 제언 **21**

References **22**

List of Tables

4.1 전체평균으로 계산한 RMSE 20

List of Figures

2.1 제한볼츠만머신의 구조 4

3.1 사용자 기반 제한볼츠만머신의 구조 10
3.2 상품 기반 제한볼츠만머신의 구조 13
3.3 하이브리드 제한볼츠만머신의 구조 16

4.1 평점자료의 히스토그램 18
4.2 제한볼츠만 머신을 이용한 추천시스템 모형 성능 19

Chapter 1

서론

사람들은 그들이 경험한 사실에 기반하여 자신만의 가치관을 만들어 살아간다. 한 사람의 경험은 다른 사람과 정확히 일치할 수 없고, 막상 같은 경험을 겪었더라도 받아들이는 가치는 다를 수 있다. 따라서 모든 사람들은 그들 고유의 특성을 가지며 살아간다. 이러한 고유성은 삶의 방식을 넘어서 구매패턴으로도 이어진다. 현재 이윤을 추구하는 여러 산업군들은 그들이 가지고 있는 자료에 기반하여 고객들에게 상품을 추천하는 시스템을 구축하고 있다. 고객마다 다른 고유한 구매패턴을 분석하고 이해해서 최적의 상품을 추천하여 매출을 올리는 것이 그들의 목표이다.

증가하는 수요에 따라 추천시스템을 위한 많은 분석방법들이 개발되어 왔다. 그 중 협업필터링(Collaborative Filtering)은 고객이 사전에 상품에 대해 평가한 자료를 이용하여, 구매한 경험이 없지만 고객이 관심 있어할 상품을 예측하여 추천해준다. 이 방법은 사용자 기반 방식과 상품 기반 방식으로 나누어 설명된다. 먼저 사용자 기반 방식이란 고객들이 구매한 상품을 기준으로 추천하는 방법으로, 구매한 상품이 가장 유사한 다른 고객을 찾아, 추천대상 고객이 구매하지 않았지만 유사고객이 구매 후 만족했던 상품을 추천해 주는 방식이다.

반면 상품 기반 방식은 상품들 간의 유사성을 먼저 파악한다. 예를 들어 상품A와 상품B는 같이 구매된 횟수가 많아 유사성이 높은 상품군이다. 이 때 특정 사용자가 상품A를 구매했고 만족도가 높았다면, 상품A와 가장 연관있는 상품B를 추천하는 방식으로 이루어진다.

협업필터링 방법은 고객들의 유사성, 상품들의 유사성을 기반으로 자동적인 상품 추천을 제공해줄 수 있다. 해당 모형이 완벽하게 구축되기 위해서는 총 N명의 사용자가 전체 M개의 상품에 대해 평가한 N×M 평점행렬이 빈틈없이 채워져 있어야 한다. 예상가능한 바와 같이 실제자료에서는 완벽한 평점행렬을 얻기는 매우 힘들다. 많은 상품 종류를 모든 사용자가 경험하고, 평가하는 것은 사실상 불가능하기 때문이다. 따라서 협업필터링을 위한 평점행렬을 완성시켜 정확한 추천을 도모하기 위해서는 추가적인 분석 방법론이 필요하다.

본 논문에서는 제한볼츠만머신(RBM:Resricted Bolrtzmann Machine)을 이용하여 협업필터링 추천방법을 개선하는 알고리즘을 소개한다. 먼저 논문의 2장에서는 제한볼츠만머신의 구조 및 학습알고리즘에 대해 살펴보고, 3장에서 협업필터링 방식을 이용한 제한볼츠만머신이 추천시스템에 사용될 수 있는 세 가지 방법론을 소개한다. 마지막으로 4장에서 한국 영화 평점 자료를 이용하여 소개한 알고리즘을 실제 분석해보고 각 방법론을 비교해 본다.

Chapter 2

제한볼츠만머신

기계학습에서 설명변수와 종속변수의 쌍(pair)으로 자료가 존재하지 않고, 종속변수만으로 구성된 자료를 분석하는 것을 비지도학습이라고 한다. 비지도학습의 한 종류로 제한볼츠만머신(RBM:Restricted Boltzmann Machines)을 예로 들수 있다. 이 모형은 최초로 Smolensky에 의해 1986에 제안되었고, 2006년 G.E. Hinton의 딥러닝 관련 논문에 언급되어 유명세를 타게 되었다. 제한볼츠만머신은 차원감소, 추천시스템 그리고 딥러닝의 기본 구조로 사용되는 등 여러가지 목적을 위해 활발히 사용되고 있다. 이 장에서는 제한볼츠만머신의 핵심 구조 및 확률분포와 모수를 학습하는 알고리즘을 파악해보려 한다.

2.1 제한볼츠만머신의 구조

제한볼츠만머신은 가시층(visible layer)과 은닉층(hidden layer) 2가지의 층으로 구성된 확률그래프모형이다. 가시층이란 실제 관측된 자료를 의미하고 은닉층이란 관측되지 않은 잠재적 특성을 의미한다. 이때 두가지 층의 모든 개체들이 0또는 1값을 갖는 이진변수만을 사용하도록 제한한다.

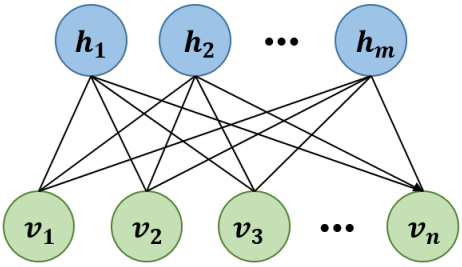

Figure 2.1: 제한볼츠만머신의 구조

Figure 2.1에는 제한볼츠만머신이 가지는 특성을 시각적으로 확인할 수 있다. 그림에서 가시층과 은닉층은 서로 모든 조합에 관해 연결되어있지만, 층 내부 개체 사이에는 어떠한 연결도 없다. 이러한 내부연결의 제약이 모형의 제한이라는 명칭으로 직결된다. 반면 모든 조합(층내부 + 층외부)에 관해 연결이 있는 모형을 볼츠만머신(BM:Boltzmann Machines)이라고 한다.

출력값은 가시층과 은닉층의 연결강도에 따라 달라진다. 연결의 강도는 가시층과 은닉층의 차원을 고려한 가중치행렬에 의해 결정되는데 이 행렬은 해당모형의 모수로 구해야하는 값이다. 머신러닝에서는 일련의 학습 알고리즘 따라 모수를 업데이트하여 그 예측치를 얻는다. 이러한 학습 알고리즘은 제한볼츠만머신이 가정하는 확률분포에 기초하여 얻어지기 때문에 알고리즘 파악에 앞서 확률분포를 파악하는 과정이 필요하다.

2.2 제한볼츠만머신의 확률분포

제한볼츠만머신은 에너지함수를 이용하여 가시층과 은닉층의 상태를 가정한다. 에너지 함수는 다음으로 정의된다.

$$\begin{aligned} E(v,h) &= -b'v - c'h - h'Wv \\ &= -\sum_{i=1}^{n} b_i v_i - \sum_{j=1}^{m} c_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} v_i h_j \end{aligned} \tag{2.1}$$

여기서 b는 가시층의 편차이고 c는 은닉층의 편차이며 W는 가중치행렬이다. 에너지 함수를 가정하고 나면, 가시층과 은닉층의 결합분포 $P(v,h)$, 은닉층이 주어진경우 가시층의 조건부분포 $P(v|h)$ 그리고 가시층이 주어진 경우 은닉층의 조건부분포 $P(h|v)$의 3가지의 확률분포를 결정할 수 있다.

1. 결합분포

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)} \qquad \text{where} \quad Z = \sum_v \sum_h e^{-E(v,h)} \tag{2.2}$$

   가시층과 은닉층의 조건부 분포는 식(2.1)의 에너지 함수를 통해 정의되고 이때 Z는 확률분포가 되도록 도와주는 정규화 상수이다.

2. 은닉층의 조건부분포

$$
\begin{aligned}
P(h|v) &= \frac{P(v,h)}{P(v)} \\
&= \frac{1}{P(v)}\frac{1}{Z}exp(c'h + b'v + h'wv) \\
&= \frac{1}{P(v)}\frac{1}{Z}exp(b'v)exp(c'h + h'wv) \\
&= \frac{1}{Z'}exp(c'h + h'wv) \\
&= \frac{1}{Z'}exp(\sum_{j=1}^{m} c'_j h_j + \sum_{j=1}^{m} h_j w_{.j} v_{.}) \\
&= \frac{1}{Z'}exp(\sum_{j=1}^{m} h_j (c_j + w_{.j} v_{.})) \\
&= \frac{1}{Z'}\prod_{j=1}^{m} exp(h_j (c_j + w_{.j} v_{.}))
\end{aligned}
$$

$$
\therefore P(h_j = 1|v) = \sigma(c_j + \sum_{i=1}^{n} w_{ij} v_i) \tag{2.3}
$$

조건부분포의 정의를 이용하여 가시층이 주어진 경우 은닉층의 조건부분포를 구할 수 있다. 이 때 σ는 시그모이드함수를 의미한다.

3. 가시층의 조건부분포

$$
P(v_i = 1|h) = \sigma(b_i + \sum_{j=1}^{m} w_{ij} h_j) \tag{2.4}
$$

2.의 방법과 동일하게 조건부분포 정의를 적용해서 쉽게 얻을 수 있다.

2.3 제한볼츠만머신의 모수 및 학습 알고리즘

해당 모형에서 모수는 가시층의 편차, 은닉층의 편차, 연결강도를 의미하는 가중치행렬로 $\theta = \{b_i, \cdots, b_n, c_i, \cdots, c_m, w_{11}, \cdots, w_{nm}\}$으로 표현된다. 주로 모

수를 구하기 위해서는 주어진 모수에 대해 가시층이 가장 잘 관측되도록 로그 우도를 가장 최대화 시키는 값들을 선정한다.

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} logP(v_i|\theta) \\ &= \sum_{i=1}^{n} \sum_{h} \frac{1}{Z} exp(-E(v_i,h)) \\ &= \sum_{i=1}^{n} log \sum_{h} exp(-E(v_i,h)) - nlogZ \\ &= \sum_{i=1}^{n} log \sum_{h} exp(-E(v_i,h)) - nlog \sum_{v,h} exp(-E(v,h)) \end{aligned} \tag{2.5}$$

식(2.5)를 θ에 관해 미분하면 다음과 같다.

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{\partial log \sum_h exp(-E(v_i,h))}{\partial \theta} - n \frac{\partial log \sum_{v,h} exp(-E(v,h))}{\partial \theta} \\ &= \sum_{i=1}^{n} \frac{\sum_h exp(-E(v_i,h)) \frac{\partial -E(v_i,h)}{\partial \theta}}{\sum_h exp(-E(v_i,h))} - n \frac{\sum_{v,h} exp(-E(v,h)) \frac{\partial -E(v,h)}{\partial \theta}}{\sum_{v,h} exp(-E(v,h))} \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \sum_{h} P(h|v_i) \frac{\partial - E(v_i,h)}{\partial \theta} - n \sum_{v,h} P(v,h) \frac{\partial - E(v,h)}{\partial \theta} \\ &= \sum_{i=1}^{n} \mathbb{E}_h \left[\frac{\partial -E(v_i,h)}{\partial \theta} | v_i \right] - n \mathbb{E}_{v,h} \left[\frac{\partial -E(v,h)}{\partial \theta} \right] \end{aligned} \tag{2.6}$$

식(2.6)의 첫번째 부분은 관측된 훈련자료와 은닉층의 조건부분포를 이용하면 쉽게 계산할 수 있다. 그러나 두번째 부분은 모든 가시층과 은닉층의 조합에 관해 계산해야 하므로 정확한 값을 도출하기 위해서 많은 시간이 걸린다. 때문에

이러한 계산시간을 효과적으로 줄이기 위해서 cd(contrastive divergence)방법이 제안되었다(Salakhutdinov et al 2007).
cd방법은 MCMC 알고리즘 중 하나인 깁스샘플링을 이용한다. 깁스샘플링을 이용하면, 다음의 과정으로 모수를 학습하게 된다.

1. 초기값설정: W_0, b_0, c_0
2. 훈련자료 v_1이용, 은닉층 도출: $p(h_{1j}|v_1) \rightarrow h_1$
3. 은닉층 h_1이용, 새로운 가시층 도출: $p(v_{2j}|h_1) \rightarrow v_2$
4. 새로운 가시층 v_2이용, 새로운 은닉층 도출: $p(h_{2j}|v_2)$
5. 업데이트
   - $W \rightarrow W + \alpha(h_1 v_1' - p(h_2|v_2)v_2')$
   - $b \rightarrow b + \alpha(h_1 - p(h_2|v_2))$
   - $c \rightarrow c + \alpha(v_1 - v_2)$
6. 2-5번 과정을 반복

깁스샘플링에서는 일반적으로 샘플을 뽑는 과정을 많이 실시하여 체인이 수렴하도록 하지만, 이 모형의 경우 반복수와 비례하여 계산시간이 증가하기 때문에 이 과정을 몇번만 실행하여 계산량을 줄인다. 알고리즘에서 α는 학습율 상수로 최적의 해를 찾기위해 한 번의 반복동안 움직이는 양을 결정한다. 이 상수가 매우 크면 최적해를 찾기위한 공간을 넘어가버리는 오버슈팅(overshooting) 현상이 발생하고, 너무 작은 경우 최적값을 찾아가기위해 오랜 시간이 걸린다는 단점이 있다.

Chapter 3

추천시스템을 위한 제한볼츠만머신

추천시스템에서 가장 많이 사용되고 있는 방법으로는 협업필터링(Collaborative Filtering)이 있다. 협업필터링의 종류에는 사용자 기반 방법과 상품 기반 방법이 있다. 먼저 사용자 기반 방법은 사용자의 과거 구매이력에 기초하여 가장 유사한 다른 사용자를 도출하는 방식이다. 예를들어 A, B, C라는 영화에 각각 10점, 9점, 8점을 준 사용자$_1$가 있을 때, B, C라는 영화에 같은 점수를 부여한 사용자$_2$가 있다. 이때, 사용자$_1$과 가장 유사한 구매 패턴을 보이는 사람이 사용자$_2$라면, 사용자$_2$에게 그와 가장 유사한 고객 사용자$_1$이 좋은 평가를 한 영화A를 추천한다. 이 방법은 가장 근접한 사용자를 찾는 것으로 최근접이웃(Nearest Neighbor Algorithm)이라고도 불린다. 반면, 상품 기반 방법은 상품들 사이의 유사성을 먼저 찾는 방법이다. 이때, 상품 간의 유사성이란 같이 구매된 횟수로 판단하여 구한다. 이 장에서는 이러한 협업필터링에 제한볼츠만머신을 활용하여 추천 알고리즘을 구현하는 방법을 소개하고 추가적으로 두가지 모형을 결합하는 하이브리드 제한볼츠만머신을 소개한다.

3.1 사용자 기반 제한볼츠만머신

분석을 위한 자료는 N명의 사용자가 M개의 상품을 1점부터 K점까지로 평가한 N×M 평점행렬이다. 평점행렬의 대부분은 평가되지 않은 빈공간으로 이루어졌기 때문에 협업필터링을 정확하게 고안할 수 없다. 따라서 제한볼츠만머신을 사용하여 협업필터링만 사용했을 때의 한계를 개선한다. 사용자 기반 제한볼츠만머신은 사용자 마다 고유의 특성이 있다는 것을 감안하여 사용자마다 다른 제한볼츠만머신을 사용한다. 이 때, 서로 다른 사용자가 같은 상품을 평가한 경우, 은닉층과 가시층의 가중치가 동일하도록 한정한다. 다시 말해, Figure 3.1에서 볼 수 있는 것 처럼 사용자 마다 다른 모형을 적용하지만 user_1과 user_2는 영화2와 영화3을 공통적으로 보았기 때문에, 해당 부분의 가시층과 은닉층이 연결되는 가중치가 동일하도록 적용하는 것이다.

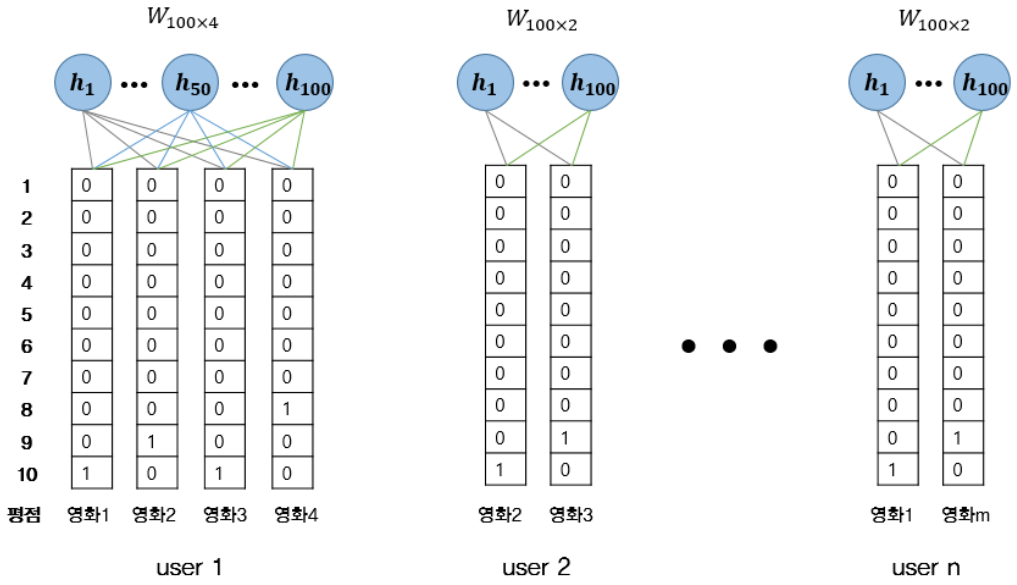

Figure 3.1: 사용자 기반 제한볼츠만머신의 구조

자료가 훈련자료, 시험자료로 나누어져 있을 경우 사용자 기반 제한볼츠만

머신은 다음과 같은 방식으로 2개의 모형을 적합한다.

- 모형1: 훈련자료의 모든 상품에 대해 개별적인 제한볼츠만머신 적용.
  - 가시층행렬:

$$V^j = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & \cdots & K \end{matrix} \\ \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \begin{array}{l} \text{사용자}_1 \\ \text{사용자}_2 \\ \text{사용자}_3 \\ \vdots \\ \text{사용자}_n \end{array}$$

  - b_j: 상품j의 가시층 편차. $(1 \times K)$
  - W_j: 상품j의 가시층과 은닉층 사이의 가중치. $(\#\text{ of 은닉층} \times K)$

- 모형2: 시험자료의 모든 사용자에 대해 훈련자료에서 평가한 자료를 이용 개별적인 제한볼츠만머신 적용.
  - 가시층행렬:

$$V^i = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & \cdots & K \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \end{array} \begin{array}{l} \text{상품}_1 \\ \text{상품}_2 \\ \text{상품}_3 \\ \vdots \\ \text{상품}_n \end{array}$$

  - c_i: 사용자i의 은닉층 편차. $(1 \times$ 은닉층갯수$)$

모형1은 같은 상품을 구매한 고객에 대해 같은 가중치를 사용하기 위해 필요하다. 모형2는 사용자 별로 서로 다른 특성을 가졌기 때문에 개인적인 고유성을 반영하기 위해 사용된다.

추천시스템의 평가를 위해서 훈련자료를 이용하여 구축한 모형을 시험자료에 적용하여 평점을 예측할 수 있다. 예측은 다음의 과정으로 이루어진다. 표기법의 혼돈을 막기위해서, i는 사용자, j는 상품, k는 점수로 한정하고 v_i^k는 선택된 사용자 한명이 상품 j에 k라는 점수를 주었다는 것을 의미한다.

1. $p(h_p = 1|V) = \sigma(c_p + \sum_{j=1}^{M}\sum_{k=1}^{K} v_j^k w_{jp}^k) \rightarrow \widehat{h_1}, \cdots, \widehat{h_P}$
   - c_p: 시험자료의 한 명의 사용자에 대해, 훈련자료에서 구축한 모형2의 은닉층 편차. 은닉층의 갯수 P만큼 존재.
   - $v_j^k w_{jp}^k$: 선택된 사용자가 j라는 상품을 k라고 평가한 경우, 모형1의 상품 j의 가중치행렬의 k번째 열.
2. $p(v_{j^*} = 1|\widehat{h_1}, \cdots, \widehat{h_P}) = \frac{exp(b_{j^*}^k + \sum_{p=1}^{P} \widehat{h_p} w_{j^*p}^k)}{\sum_{s=1}^{K} exp(b_{j^*}^s + \sum_{p=1}^{P} \widehat{h_p} w_{j^*p}^s)}$
   - b_{j^*}: 시험자료에 새로운 상품이 j^*일 때, 모형1의 상품 j^*에 대한 가시층 편차.
   - $w_{j^*p}^k$: 모형1의 새로운 상품 j^*에 대한 가중치행렬의 k번째 열.

첫 번째 과정에서는 시험자료에 사용자의 과거 상품 이용자료를 통해 잠재된 은닉상태를 추정한다. 두번째 과정은 첫번째 과정에서 추정된 은닉상태를 이용하여 시험자료의 새 상품에 대해 해당 사용자가 어떤 평가를 내릴지 예측한다.

3.2 상품 기반 제한볼츠만머신

상품 기반 제한볼츠만머신은 사용자 기반 제한볼츠만머신이 사용자별로 다른 모형을 사용한 것에 반해, 상품 별로 상이한 모형을 구축하여 적용한다. Figure 3.2의 예시처럼 같은 영화를 본 사용자들의 점수 자료를 가시층에 두고 각 영화모형 별로 동일하게 100개의 은닉층을 가정하고, 가중치와 편차를 학습한다. 이는 Figure 3.1이 사용자 별로 평가한 영화를 입력으로 하여 모형을 구축한 방식에서 주체를 변경한 모형이다.

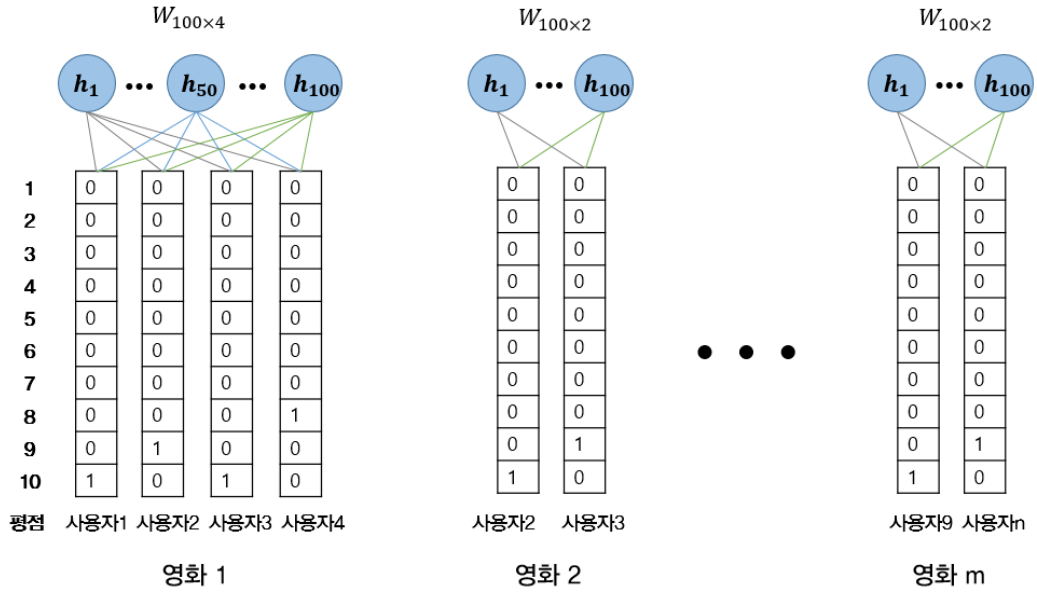

Figure 3.2: 상품 기반 제한볼츠만머신의 구조

이 모형에서도 서로다른 영화를 같은 사용자가 본 경우, 가시층과 은닉층의 가중치가 같도록 조절하기 위해 다음의 방식으로 모형을 적합한다.

- 모형1: 훈련자료의 모든 사용자에 대해 개별적인 제한볼츠만머신 적용.
  - 가시층행렬:

$$V^i = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & \cdots & K \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \end{array} \begin{matrix} \text{상품}_1 \\ \text{상품}_2 \\ \text{상품}_3 \\ \vdots \\ \text{상품}_n \end{matrix}$$

  - b_i: 사용자i의 가시층 편차. $(1 \times K)$
  - W_i: 사용자i의 가시층과 은닉층 사이의 가중치. $(\#\text{ of 은닉층} \times K)$
- 모형2: 시험자료의 모든 상품에 대해 훈련자료에서 평가한 자료를 이용 개별적인 제한볼츠만머신 적용.
  - 가시층행렬:

$$V^j = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & \cdots & K \end{matrix} \\ \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \begin{matrix} \text{사용자}_1 \\ \text{사용자}_2 \\ \text{사용자}_3 \\ \vdots \\ \text{사용자}_n \end{matrix}$$

  - c_i: 상품j의 은닉층 편차. $(1 \times$ 은닉층갯수$)$

이 방법에서 모형1은 같은 사용자를 공유한 상품에 대해 같은 가중치를 사용하기 위해 필요하고 모형2는 상품의 특성을 반영하기 위해 사용된다. 그리고 예측은 사용자 기반 제한볼츠만머신과 동일하게 두가지 단계를 통해 이루어진다.

여기서 v_i^k는 선택된 상품에대해, 사용자i가 점수 k를 부여했다는 것을 의미한다.

1. $p(h_q = 1|V) = \sigma(c_q + \sum_{i=1}^{N}\sum_{k=1}^{K} v_i^k w_{iq}^k) \quad \rightarrow \widehat{h_1}, \cdots, \widehat{h_Q}$
   - c_q: 시험자료의 하나의 상품에 대해, 훈련자료에서 구축한 모형2의 은닉층 편차. 은닉층의 갯수 Q만큼 존재.
   - $v_i^k w_{iq}^k$: 선택된 상품이 i라는 사용자에 대해 k라고 평가된 경우, 모형1의 사용자 i의 가중치행렬의 k번째 열.
2. $p(v_{i^*} = 1|\widehat{h_1}, \cdots, \widehat{h_Q}) = \frac{exp(b_{i^*}^k + \sum_{q=1}^{Q} \widehat{h_q} w_{i^*q}^k)}{\sum_{s=1}^{K} exp(b_{i^*}^s + \sum_{q=1}^{Q} \widehat{h_q} w_{i^*q}^s)}$
   - b_{i^*}: 시험자료에 새로운 사용자가 i^*일 때, 모형1의 사용자 i^*에 대한 가시층 편차.
   - $w_{i^*q}^k$: 모형1의 새로운 사용자 i^*에 대한 가중치행렬의 k번째 열.

3.3 하이브리드 제한볼츠만머신

하이브리드 제한볼츠만머신은 앞의 두가지 방법을 혼합하여 사용하는 모형이다. 특별한 방법을 사용하여 새롭게 학습하는 알고리즘이 아니라, 이미 학습을 끝낸 사용자 기반 제한볼츠만머신과 상품기반 제한볼츠만머신의 모수들을 이용한다. Figure 3.3에서 그 구조를 파악할 수 있는데, 이 때 가시층의 하나의 개체는 사용자i가 상품j를 평가한 점수를 의미한다. 그러한 가시층 개체는 사용자 기반 모형에서 사용자i에 대해서만 모형이 구축되므로 사용자i가 평가한 상품들에게만 연결되어 모든 조합에 관해 은닉층과 연결되어 가중치 행렬을 만든다. 반면 상품 기반 모형에서는 상품j를 주체로 하여, 상품j를 평가한 사용자들에게만 연결되어 모수를 학습한다.

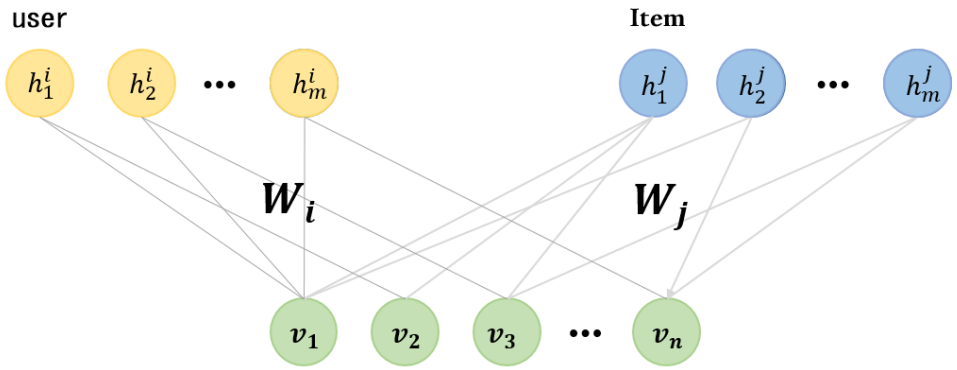

Figure 3.3: 하이브리드 제한볼츠만머신의 구조

예측은 새롭게 모형을 적용하여 훈련시키는 것이 아니라, 기존에 학습시킨 사용자 기반 방법과 상품 기반 방법에서 예측된 평가점수 값의 평균을 하이브리드 제한볼츠만머신의 예측 평가값으로 사용한다.

Chapter 4

시뮬레이션

4.1 데이터 설명

소개한 추천시스템방법들을 적용하기 위해 사용한 데이터는 CGV 고객-영화 평점 데이터이다. 해당 자료는 크롤링 방법을 통해 얻을 수 있었다. 데이터는 2013년 1월 부터 12월 까지 1년 동안 얻어진 912,407개의 평점 자료로, 고객 96,979명이 총 800편의 영화를 평가했다. 모든 고객과 모든 영화의 조합을 고려했을 때, 1.18% 채워져 있는 자료로 행렬의 희소성 문제가 상당하다. 점수는 0점 부터 10점까지 11개의 지표로 얻어져있고, Figure 4.1에서 그 분포를 파악할 수 있다.

각 모형들의 평가를 위해서 사용한 지표는 RMSE(Root Mean Square Error)이다. 이 지표는 n개의 시험자료에 대해, 고객-영화 조합에 대한 예측된 평점과 실제 평점의 차이를 제곱한 값들의 평균값을 의미한다. 해당 지표는 각 모형이 가질수 있는 오차에 대한 척도로 그 값이 작을수록 모형성능이 더 좋다고 판단할 수 있다.

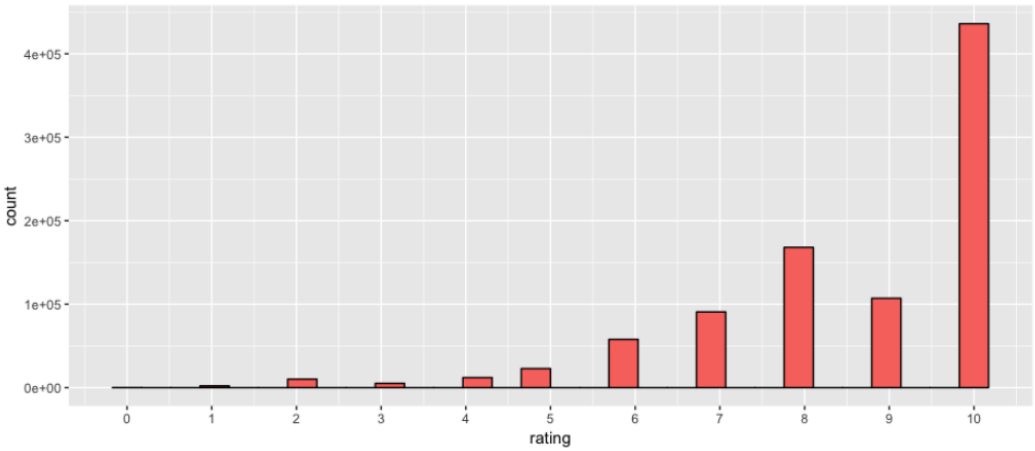

Figure 4.1: 평점자료의 히스토그램

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \widehat{r}_i)^2} \tag{4.1}$$

자료는 제한볼츠만머신을 학습하기 위한 훈련자료(911,907개)와 머신의 평가를 위한 시험자료(500개)로 나누었다. 나눈 기준으로는 시험자료의 고객이 훈련자료의 부분집합이 되어 훈련자료에서 만든 사용자 별(혹은 상품 별) 영화 평점 모형이 시험자료에 사용될 수 있도록 설정하였다. 학습을 위해 미리 지정해야할 은닉층의 갯수는 사용자 기반, 상품 기반의 방법에서 모두 50개로 하였다. 추가적으로 학습율 상수는 Georgiev at al(2013).을 따라 사용자 모형의 경우는 $\frac{0.05}{\text{훈련자료의 사용자 수}}$, 상품 모형에서는 $\frac{0.05}{\text{훈련자료의 영화 수}}$로 지정하여 학습시켰다. 훈련자료의 사용자 수가 영화 수에 비해 월등히 많기 때문에 상품 기반의 모형에서 학습율이 더 작다. 또 cd알고리즘을 설정하기 위해서는 각 반복 당 실시한 표본추출의 갯수를 설정하여야 한다. 횟수가 많을 수록 정확도가 증가하지만 계산량에 따른 시간도 증가하기 때문에 2로 한정하였다.

추가적으로 입력해야하는 것으로 epoch의 크기가 있다. epoch이란 훈련자료를 학습에 몇번 활용할지를 결정하는 상수이다. 예를들어, epoch이 1인 경우 훈련자료를 학습에 한번만 사용하는 것이고, 그 수가 커질수록 훈련자료를 학습에 많이 이용한다. epoch의 크기와 오버피팅(overfitting)문제는 양날의 검과 같아서 모형들을 학습할때 마다 손실함수를 확인하므로 적절한 수를 결정하는 것이 바람직하다.

입력되어야할 모든 값들을 적용하여 CGV 고객-영화 평점 데이터를 분석하였다. 소개한 세가지의 모형(사용자 기반 방식, 상품 기반 방식, 하이브리드 방식)을 epoch크기를 증가시키며 학습하였고 그 결과는 Figure4.2를 통해 확인할 수 있다.

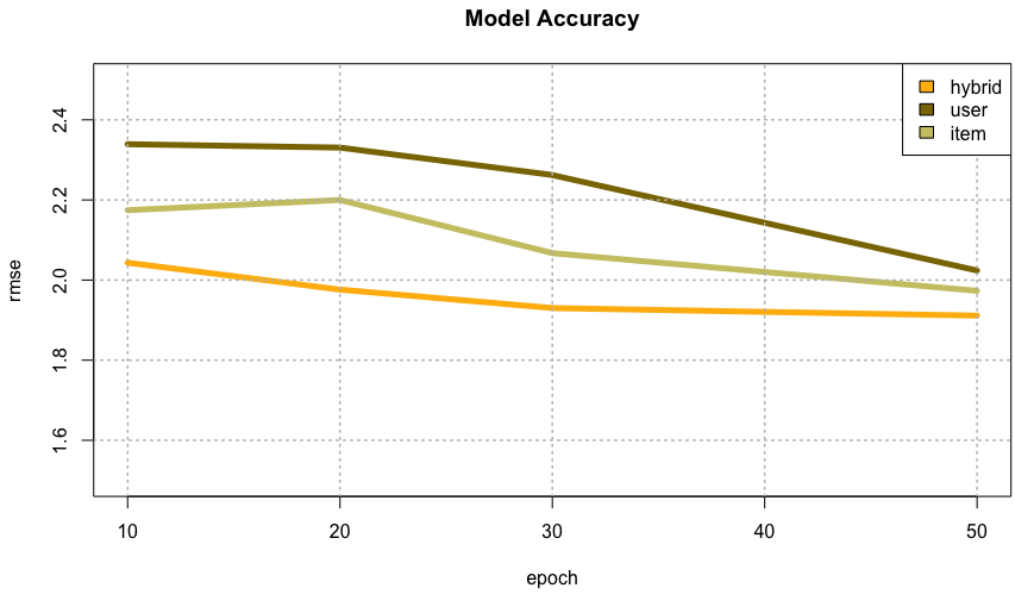

Figure 4.2: 제한볼츠만 머신을 이용한 추천시스템 모형 성능

학습 결과를 보면 세가지 모형 모두에서 모형 정확성의 척도로 사용

한 RMSE가 epoch크기의 증가와 함께 줄어드는 것을 확인 할 수 있었다. 이는 epoch의 크기가 모형 정확도에 양의 방향으로 영향을 끼친다는 것을 보여준다. 개별 모형들을 비교해 보면, 세가지 모형 중 사용자 기반 방법이 예측력이 가장 떨어졌고, 하이브리드 모형이 가장 좋은 예측력을 보인다는 것을 확인 할 수 있다. 사용자 기반 모형의 경우 96,979명의 고객에 대해 서로 다른 제한볼츠만머신을 적용했기 때문에, 학습에 필요한 시간이 상당했다. 반면 상품 기반 모형의 경우 800편의 영화마다 다른 모형들을 적용하였기 때문에 상대적으로 빠른시간 안에 학습된 결과를 확인 할 수 있었다. 두가지 모형을 비교해 보면, 개별 고객들이 영화 평점에 미치는 영향력 보다, 하나의 영화에 관해 사람들이 평가한 점수가 평점예측에는 더욱 효율적이며 정확했다. 그리고 고객중심과 상품중심의 모형학습 결과를 평균한 하이브리드 모형의 경우, 한가지 모형이 극단적으로 예측할 수 있는 확률을 다른 모형과의 결합을 통해 낮추었다.

구분	전체	사용자	상품
RMSE	3.2954	2.651	2.729

Table 4.1: 전체평균으로 계산한 RMSE

제한볼츠만머신의 효과를 확인하기 위해, 전체 평균 값으로 시험자료의 영화평점을 예측한 결과와 비교를 실시하였다. Table 4.1에는 훈련자료의 전체 평균으로 시험자료를 예측한 결과, 사용자 별 평점평균으로 시험자료를 예측한 결과 그리고 영화 별 평점평균으로 시험자료를 예측한 결과가 제시되어있다. 제시된 3가지 값에서 모두 본 논문에서 제안한 모형보다 낮은 성능을 보인다. 평균으로는 고객이 전반적으로 영화를 보고 느끼는 만족도패턴이나, 영화 자체의 인기유무를 확인할 수 있지만 정확한 예측에는 한계가 있다는 부분을 시사하는 결과이다.

Chapter 5

결론 및 제언

추천시스템은 존재하는 많은 마케팅 기법과 함께 기업으로 하여금 새로운 방식으로 고객들을 유치할 수 있도록 돕는다. 본 논문에서는 그중 제한볼츠만 머신을 이용하여 추천시스템을 구현하는 방법을 제시하였다. 제한볼츠만머신은 머신러닝기법 중 비지도학습의 한가지 종류로, 특별한 입력변수가 존재하지 않는 단일자료의 경우 사용하는 모형이다. 대표적인 추천시스템 방법 중 하나인 협업필터링 방법에 해당 머신을 적용하는 알고리즘을 총 3가지 방법으로 제시하였다. 첫째, 사용자 기반 방법은 사용자 별로 상품을 평가하는 상이한 특성을 사용자 별로 다른 모형을 만드는 것으로 반영하였다. 둘째, 상품 기반 방식은 하나의 상품에 대한 고객들의 공통된 의견을 반영하는 방법이다. 셋째, 하이브리드 모형은 사용자 기반 방식과 상품기반 방식의 예측값을 평균한 것이다. 이 방식은 앙상블 모형에서 배깅(Bagging)이 약한 모형들을 평균으로 결합하여 예측력을 증가하시키는 방식과 유사하다.

자료분석에 추가적으로 영화에 대한 정보나 사용자에 관한 정보가 입력정보로 제시되는 경우 다른 지도학습기법이나 딥러닝 기법으로 예측력을 더욱 상승시킬 여지가 있다.

References

Georgiev, Kostadin, and Preslav Nakov. "A non-iid framework for collaborative filtering with restricted boltzmann machines." International Conference on Machine Learning. 2013.

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, (2012): 983–993.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, (2001): 638–648.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

Fischer, Asja, and Christian Igel. "An introduction to restricted Boltzmann machines." Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (2012): 14-36.

Abstract

Jeong Ba Raem
The Department of Statistics
The Graduate School
Seoul National University

Recommender systems initially recommended items for users using overall sales pattern. However, the system that depended on the global mean is ineffective because it doesn't reflect individual personalities and product characteristics. In order to solve this problem, the collaborative filtering (CF) method was used. CF suggests a new product to the user based on either User based CF or Item based CF. However, the accuracy of the recommendation using CF can be lowered by the sparsity of the user item matrix. In this case, Restricted Boltzmann Machine, which is one of the unsupervised learning methods, can be used to improve the quality of the recommendations. This paper introduces a new collaborative filtering based on Restricted Boltzmann Machine and apply it to actual Korean movie rating data.

Keywords : Restricted Boltzmann Machine, Unsupervised learning, , Collaborative Filtering, User based CF, Contents based CF

Student Number : 2016-20275