

## 커널 함수를 도입한 새로운 추천 시스템

### A New Kernelized Approach to Recommender System

이제현\* · 황재필\*\* · 김은태\*\*

Jaehun Lee, Jae Pil Hwang, and Euntai Kim

\* 연세대학교 전기전자공학부

\*\* 현대자동차그룹

#### 요약

본 논문에서는 커널 함수를 이용한 기법을 통한 추천 시스템을 제안한다. 제안된 추천 시스템은 기계 학습 기법을 이용하여 새로운 아이템에 대한 사용자의 선호도를 예측하고 예측된 결과를 바탕으로 사용자가 선호할만한 아이템들을 추천한다. 일반적으로 사용자의 평가 정보는 잡음이 포함되어 있고 일관성이 적으므로 잡음에 영향을 적게 받는 이원 분류기인 이중 마진 Lagrangian support vector machine (DMLSVM) 을 사용한다. 제안된 기법은 MovieLens 데이터베이스에 적용하였다. 또한 시뮬레이션을 통해 제안된 방법의 우수성을 확인하였다.

**키워드** : 추천 시스템, 이중 마진 Lagrangian support vector machine, MovieLens

#### Abstract

In this paper, a new kernelized approach for use in a recommender system (RS) is proposed. Using a machine learning technique, the proposed method predicts the user's preferences for unknown items and recommends items which are likely to be preferred by the user. Since the ratings of the users are generally inconsistent and noisy, a robust binary classifier called a dual margin Lagrangian support vector machine (DMLSVM) is employed to suppress the noise. The proposed method is applied to MovieLens databases, and its effectiveness is demonstrated via simulations.

**Key Words**: Recommender system, Dual margin Lagrangian SVM, MovieLens

## 1. 서론

최근 웹의 정보량이 폭발적으로 증가함에 따라 추천 시스템 (recommender system) 이 도서, 음악, 영화 등의 다양한 종류의 온라인 사업자들에게 관심을 받고 있다. 추천 시스템은 전문가 시스템의 일종으로 사용자들의 과거의 선호도를 바탕으로 사용자에게 적절한 아이템을 추천해주는 것을 목적으로 한다. 몇몇 추천 시스템의 적용 예가 연구되어 왔는데, 대표적인 예로 도서와 음반을 추천하는 Amazon.com [1], 영화를 추천해주는 MovieLens [2], 그리고 뉴스를 추천해주는 VERSIFI Technologies [3] 등이 있다.

추천 시스템에 대한 연구는 내용 기반 추천, 협업적 여과, 혼합 방식의 3가지로 크게 분류할 수 있다. 내용 기반 추천 알고리즘은 새로운 아이템에 대한 사용자의 선호도를 해당 사용자의 다른 아이템들에 대한 기존의 평가에 기반하여 예측한다. 내용 기반 추천 방법에서는 사용자가 과거에 선호하던 아이템과 비슷한 아이템들이 추천목록에 포함된다 [4]. 협업적 여과 방법은 사용자의 평점을 사용자와 비슷한 성향의 다른 사용자들의 평점을 바탕으로 예측한다. 협업적 여과 방법

에서는 사용자와 비슷한 사용자들이 선호했던 아이템들이 추천 목록에 포함된다 [5]. 혼합 방식은 내용 기반 추천 방법과 협업적 여과 방법을 혼합하여 아이템의 평점을 예측한다.

본 논문에서는 새로운 방법의 추천 시스템을 제안한다. 제안된 방법에서는 먼저 추천 시스템을 분류 문제로 문제화한다. 이후 커널 함수를 사용하는 기계 학습 기법을 사용하여 미지의 아이템들에 대한 사용자의 선호도를 예측한다. 사용자의 평점 정보들이 일반적으로 일관성이 적고 잡음이 포함되어 있다는 점을 고려하여서 제안된 방법은 잡음에 영향을 적게 받는 이원 분류기인 이중 마진 Lagrangian 서포트 벡터 머신 (DMLSVM: Dual Margin Lagrangian Support Vector Machine) 을 사용하여 효율적이고 안정적인 추천을 한다. 논문의 구성은 다음과 같은 순서로 되어 있다. 제 2장에서는 추천 문제를 분류 문제로 공식화하고 제안하는 방법과 관련된 몇몇 기본 원리들에 대하여 기술한다. 제 3장에서는 이중 마진 Lagrangian 서포트 벡터 머신에 기반하는 추천 시스템을 제안하고 이에 대해 기술한다. 제 4장에서는 시뮬레이션과 그 결과를 요약 정리하고 마지막으로 제 5장에서는 결론으로 마무리한다.

접수일자 : 2011년 6월 23일

완료일자 : 2011년 10월 10일

+ 교신 저자

## 2. 배경지식

### 2.1 추천 문제의 분류 문제로의 전환

$N$ 개의 아이템들  $I = \{i_1, i_2, \dots, i_N\}$  과 다양한 사용자들  $u \in U$ 가 시장에 존재한다고 가정해보자. 이때  $I$ 와  $U$ 는 각각 아이템과 사용자 집합을 나타낸다. 한 사용자  $u$ 의 특정 아이템  $i_i$ 에 대한 평점  $r_i \in R$ 은 다음과 같이 표현할 수 있다.

$$r_i = f(u, i_i) \quad (1)$$

여기서  $R$ 은 가능한 평점들의 집합을 나타낸다. 예를 들어, MovieLens [2] 데이터베이스에서  $R = \{1, 2, 3, 4, 5\}$ 로 여기서 평점 1은 가장 낮은 평점이고 평점 5는 가장 높은 평점이다. 전체 아이템 집합에서 일부 아이템들  $i_j \in I_D$ 의 경우 사용자가 직접 평가를 하여서 그 평점 정보  $r_j$ 가 미리 알려져 있는 반면, 그 밖의 다른 아이템들  $i_k \in I_{ID}$ 에 대한 평가는 미지의 상태라고 가정한다. 여기서  $I_D$ 와  $I_{ID}$ 는 각각 사용자에 의해 직접 평가된 아이템 집합과 그렇지 않은 아이템들의 집합을 나타내고 전체 아이템 집합은 두 집합의 합집합으로 표현 가능하다 ( $I_D \cup I_{ID} = I$ ). 이때, 추천 시스템은 다음과 같이 정의할 수 있다.

$$\begin{aligned} \text{Predict } r_k &= f(u, i_k) \\ \text{based on } r_j &= f(u, i_j) \end{aligned} \quad (2)$$

여기서  $i_j$ 와  $i_k$ 는 비슷한 아이템을 의미한다.

이 문제를 해결하기 위해서 다양한 연구가 이루어져왔다. 예를 들어, 정보 검색 [6] 혹은 정보 여과 [7] 등의 기법을 이용한 연구가 있었고 대부분의 기존의 방법들은 유사도 함수를 이용하여서 문제를 해결하였다. 유사도 함수란 미지의 아이템  $i_k$ 에 대한 사용자의 평점을 사용자와 해당 아이템간의 유사도를 근거로 예측하는 것으로 대표적인 예로 코사인 유사도를 들 수 있다 [6].

$$\begin{aligned} r_i &= f(u, i_i) = \cos(\vec{w}_u, \vec{w}_i) \\ &= \frac{\vec{w}_u \cdot \vec{w}_i}{\|\vec{w}_u\| \times \|\vec{w}_i\|} \end{aligned} \quad (3)$$

여기서  $\vec{w}_u$ 와  $\vec{w}_i$ 는 각각 사용자와  $i$ 번째 아이템의 가중치 벡터이다. 이외에도 모델에 기반한 기법들을 이용한 추천 시스템도 있는데, 대표적인 예로 Pazzani 와 Billsus 가 제안한 naive Bayesian classifier를 이용한 추천 시스템이 있다 [8].

본 논문에서는 추천 문제를 다음과 같은 가정을 통하여 보다 단순화시켰다; (a) 각각의 아이템  $i_i \in I$ 이 특징 벡터  $\mathbf{x}_i \in R^M$ 로 표현 가능하다. (b) 각각의 아이템  $i_i \in I$ 에 이진 라벨  $y_i \in \{1, -1\}$ 를 붙인다. 여기서  $M$ 은 아이템의 특징들의 수를 나타내고 이진 라벨  $y_i$ 에서 1과 -1은 각각 “추천한다”와 “추천하지 않는다”를 나타낸다. 그렇다면 앞서 설명한 추천 문제 (2)는 다음과 같은 이진 분류 문제로 공식화할 수 있다.

$$\begin{aligned} \text{Classify } \mathbf{x}_k \\ \text{based on } (\mathbf{x}_j, y_j) \end{aligned} \quad (4)$$

위의 (4)식은 전형적인 이진 분류 문제로 볼 수 있다.

### 2.2 분류 문제의 해결

서포트 벡터 머신 (SVM: Support Vector Machine)은 대표적인 분류기중 하나로 패턴 인식 [9], 얼굴 검출 [10], 그리고 문자 인식 [11] 등등의 다양한 분야에서 사용되고 있다. SVM의 목적은 다음과 같은 초평면을 찾는 것이다.

$$y = \mathbf{x}^T \mathbf{w} + b \quad (5)$$

이 때,  $\mathbf{x}, \mathbf{w} \in R^M$ 이고  $y \in \{1, -1\}$ 이며 결정 초평면은 주어진 두 부류들 사이의 마진을 최대화시킨다. 본 논문에서는 평이한 SVM이 아닌 DMLSVM을 적용하였다 [12]. 그 이유는 추천 시스템의 주요 특징중 하나인 트레이닝 데이터로 사용되는 사용자의 평점 정보가 일관성이 적고 잡음이 있을 가능성이 높기 때문이다. DMLSVM의 기본 아이디어는 그림 1과 같다. DMLSVM에서는 특징 공간을 marginal 구간과 surplus 구간으로 구분하고 두 구간 모두에 각각 다른 가중치 벡터를 부여한다.

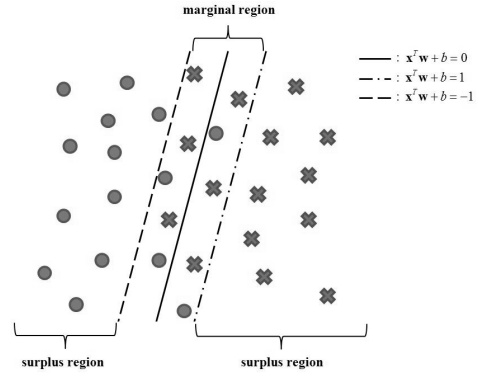


그림 1. DMLSVM의 기본 아이디어.  
Fig. 1. Basic idea of DMLSVM.

보통의 SVM과는 달리 DMLSVM에서는 모든 데이터 값들이 두 가지 가중치 혹은 값에 의해 제한을 받게 된다. 이 때 어느 가중치에 의해 제한을 받을지는 해당 데이터 값이 marginal 구간에 속하는지 혹은 surplus 구간에 속하는지에 따라 결정된다 [12]. 트레이닝 데이터의 모든 데이터 값들이 분류기에 사용되므로 DMLSVM은 트레이닝 데이터가 적은 환경에서도 보다 안정적이고 효율적인 성능을 보이며, 따라서 추천 시스템에 사용되기 적합하다.

## 3. DMLSVM을 이용한 추천 시스템

제안하는 방법의 기본 아이디어는 커널 함수를 사용한 DMLSVM을 이용하여 추천 문제를 해결한다는 것이다. 다음과 같이 표현되는 추천 시스템을 생각해보자.

$$f: U \times I \rightarrow R \quad (6)$$

여기서  $U, I, R$ 은 각각 사용자, 아이템, 평점정보의 집합을 나타낸다. 이를 이진분류문제로 전환시 다음과 같다.

$$f: U \times I \rightarrow \{1, -1\} \quad (7)$$

위 식과 같이 추천 문제를 이진 분류 문제로 바꿀 때 적극적 추천 시스템 (positive RS), 소극적 추천 시스템 (negative RS), 그리고 혼합 추천 시스템 (combined RS) 등의 3가지 종류의 방법을 제안한다. 적극적 추천 시스템은 뚜렷하게 매력적인 아이템들에게 “추천한다”는 라벨을 붙이고 나머지 아

이템들은 “추천하지 않는다”는 라벨을 붙인다. 그리고 DMLSVM은 뚜렷하게 매력적인 아이тем들을 학습하는데 초점을 맞춘다. 반면에 소극적 추천 시스템 방법에서는 확연하게 매력적이지 않은 아이тем들에게 “추천하지 않는다”는 라벨을 붙이고 나머지 아이тем들은 “추천한다”는 라벨을 붙인다. 그리고 DMLSVM은 뚜렷하게 매력적이지 않은 아이тем들을 학습하는데 초점을 맞춘다. 예를 들어,  $R = \{1, 2, 3, 4, 5\}$  일 때 평점들은 그림 2와 같이 해석한다.

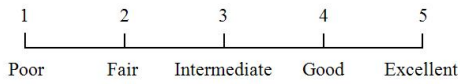


그림 2  $R = \{1, 2, 3, 4, 5\}$  일 때 평점들의 해석.

Fig. 2. Interpreting of ratings when  $R = \{1, 2, 3, 4, 5\}$ .

위의 경우 적극적 추천 시스템 방법에서는 평점 4와 5를 받은 아이тем들만이 “추천한다”는 라벨에 속하게 되고 나머지 평점 1, 2, 3을 받은 아이тем들은 “추천하지 않는다”는 라벨에 속한다. 반면에 소극적 추천 시스템 방법에서는 평점 1과 2를 받은 아이тем들만이 “추천하지 않는다” 라벨에 속하게 되고 나머지 평점 3, 4, 5를 받은 아이тем들은 “추천한다” 라벨에 속하게 된다. 마지막으로 혼합 추천 시스템 방법은 적극적 추천 시스템과 소극적 추천 시스템의 결과 모두를 이용한다. 만약 둘의 결과가 “추천한다” 이든 “추천하지 않는다” 이든 일치한다면 해당 라벨을 결과로 사용한다. 만약 둘의 결과가 일치하지 않는다면 적극적과 소극적 추천 시스템 중 보다 확신적인 시스템의 결과값을 취하게 된다.

### 3.1 적극적 추천 시스템

트레이닝 아이тем 집합  $I_D = \{i_1, i_2, \dots, i_{N_D}\}$  가 주어졌을 때 사용자  $u$ 의 각각의 아이тем들에 대한 평점이 아래와 같이 주어졌다고 가정해보자.

$$r_1 = f(u, i_1) \quad (8)$$

⋮

$$r_{N_D} = f(u, i_{N_D})$$

여기서  $i_i$ 는 특징 벡터  $\mathbf{x}_i \in R^M$ 로 표현 가능하다. 추천 라벨  $y_i$ 는 평점 정보  $r_i$ 에 따라 다음과 같이 할당된다.

$$y_i = \begin{cases} 1 & r_i > \theta_p \\ -1 & \text{otherwise} \end{cases} \quad (9)$$

즉,  $\theta_p$ 보다 큰 평점을 갖는 아이тем은 뚜렷하게 매력적인 아이тем들로 “추천한다” 라벨을 붙이고 그렇지 않은 아이тем들에는 “추천하지 않는다” 라벨을 붙인다. 이 트레이닝 데이터셋에 DMLSVM을 적용하여 가중치 벡터  $\mathbf{w}_p$ 와 bias값  $b_p$ 를 구하면 [12], 미지의 아이тем  $i_k \in I_{ID}$ 에 대한 추천여부는 다음 식을 통해 결정한다.

$$y_k = \begin{cases} 1 & \mathbf{w}_p^T \mathbf{x}_k + b_p > 0 \\ -1 & \text{otherwise} \end{cases} \quad (10)$$

또한 커널 함수를 사용한다면 추천결과는 다음과 같다.

$$y_k = \begin{cases} 1 & \mathbf{w}_p^T \phi(\mathbf{x}_k) + b_p > 0 \\ -1 & \text{otherwise} \end{cases} \quad (11)$$

여기서  $\phi()$ 는 비선형 사상 함수를 나타낸다. 예를 들어  $\phi(\mathbf{x}) = \mathbf{x}$ 는 선형 커널 함수이다.

### 3.2 소극적 추천 시스템

트레이닝 아이тем 집합  $I_D = \{i_1, i_2, \dots, i_{N_D}\}$  가 주어졌을 때 사용자  $u$ 의 각각의 아이тем들에 대한 평점이 식(8)과 같다고 할 때, 추천 라벨  $y_i$ 는 평점 정보  $r_i$ 에 따라 다음과 같이 할당된다.

$$y_i = \begin{cases} -1 & r_i < \theta_n \\ 1 & \text{otherwise} \end{cases} \quad (12)$$

즉,  $\theta_n$ 보다 작은 평점을 갖는 아이тем들은 매우 매력적이지 않은 아이тем들로 “추천하지 않는다” 라벨을 붙이고 그렇지 않은 아이тем들은 모두 “추천한다” 라벨을 붙인다. 적극적 추천 시스템에서와 같이 DMLSVM을 적용할 경우 [12], 미지의 아이тем  $i_k \in I_{ID}$ 에 대한 추천여부는 다음 식을 통해 결정한다.

$$y_k = \begin{cases} -1 & \mathbf{w}_n^T \mathbf{x}_k + b_n > 0 \\ 1 & \text{otherwise} \end{cases} \quad (13)$$

또한 커널 함수를 사용한다면 추천결과는 다음과 같다.

$$y_k = \begin{cases} -1 & \mathbf{w}_n^T \phi(\mathbf{x}_k) + b_n > 0 \\ 1 & \text{otherwise} \end{cases} \quad (14)$$

여기서  $\phi()$ 는 비선형 사상 함수를 나타낸다.

### 3.3 혼합 추천 시스템

적극적 추천 시스템과 소극적 추천 시스템이 각각 분리되어서 구축되었다고 생각해보자. 이 때 두 시스템의 결과는 모두 혼합 추천 시스템에 사용된다. 혼합 추천 시스템을 간단히 도식화하면 그림 3과 같다.

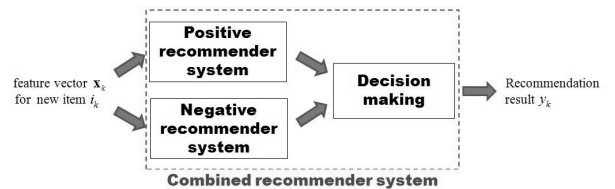


그림 3. 혼합 추천 시스템 도해.

Fig. 3. Combined recommender system.

미지의 아이тем  $i_k \in I_{ID}$ 에 대한 적극적 추천 시스템과 소극적 추천 시스템의 추천 결과를 각각  $y_k^p$ 와  $y_k^n$ 라고 정의하자. 이 때, 혼합 추천 시스템의 추천 결과  $y_k$ 는 다음과 같다.

$$y_k = \begin{cases} y_k^p & |\mathbf{w}_p^T \mathbf{x}_k + b_p| > |\mathbf{w}_n^T \mathbf{x}_k + b_n| \\ y_k^n & \text{otherwise} \end{cases} \quad (15)$$

여기서  $(\mathbf{w}_p, b_p)$ 와  $(\mathbf{w}_n, b_n)$ 은 각각 적극적 추천 시스템과 소극적 추천 시스템의 파라미터들이다. 즉, 적극적과 소극적 추천 시스템의 추천 결과가 일치할 경우, 혼합 추천 시스템의 추천 결과도 이와 동일한 결과 값을 나타낸다 ( $y_k^p = y_k^n = y_k$ ). 그렇지 않은 경우, 즉, 적극적과 소극적 추천 시스템의 추천 결과가 다른 경우 ( $y_k^p \neq y_k^n$ ), 혼합 추천 시스템은 두 추천 시스템의 결과중 보다 높은 마진을 갖는 추천 시스템의 결과를 채택한다.

### 4. 실험 및 결과 분석

제안하는 추천 시스템들은 널리 쓰이는 데이터베이스 중 하나인 MovieLens [2] 데이터베이스에 적용하여 실험이 이루어졌다. MovieLens 데이터베이스는 영화를 추천하는 웹사이트로 가장 유명한 추천 시스템 데이터베이스 중 하나이다. 여기서 데이터셋은 다양한 사용자들의 다양한 영화에 대한 평점 정보들로 이루어져 있고 평점을 취득한 기간은 1997년 9월부터 1998년 4월까지이다. 데이터셋의 주요한 특징들은 표1과 같다.

표 1. MovieLens 데이터셋의 특징.

Table 1. The characteristics of MovieLens dataset.

Parameter	Value
Number of users	943
Number of movies	1682
Number of ratings	100000
Value of ratings	{1,2,3,4,5}

데이터셋은 943명의 사용자들이 1,682개의 영화에 대해 평균 총 100,000개의 평점 정보들로 구성되어 있고, 이는 다음과 같은 구조의 텍스트 파일로 저장되어 있다.

표 2. MovieLens 데이터베이스의 구성.

Table 2. The structure of MovieLens dataset.

userID	itemID	rating	TimeStamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
⋮			

여기서 userID와 itemID는 각각 사용자와 영화의 고유번호를 나타내고 rating은 그 사용자의 해당 영화에 대한 평점 정보를 나타낸다. Timestamp는 1970/1/1 이후 몇 초의 시간이 경과했을 때 평점 정보가 저장되었는지를 나타낸다. 각각의 영화들에 대한 정보는 표3과 구조의 데이터베이스로 이루어져 있다.

표 3. MovieLens 데이터베이스의 영화 정보.

Table 3. Information of movies in MovieLens dataset.

movieID	Title	Release date	Genre
1	Toy Story	1995	animation, comedy
2	Golden Eye	1995	action, adventure, thriller
⋮			

실험을 위하여 100,000개의 데이터셋을 트레이닝 데이터와

테스트 데이터로 나누었다. 트레이닝과 테스트 데이터를 나눌 때 기준은 각 사용자마다 10개의 평점 정보가 테스트 데이터에 포함되도록 하였다. 추천 시스템에서 아이템들의 특징을 나타내는 feature로는 표3의 release date 및 genre를 사용하였다. 트레이닝과 테스트 데이터의 구성은 표4와 같다.

표 4. 실험에 사용된 데이터셋.

Table 4. The dataset used in the simulations.

Data	Number of ratings	Number of users	Number of movies
Training set	90570	943	1628
Test set	9430	943	10 per user

제안하는 추천 시스템의 성능 평가는 다음과 같은 방식을 통하여 이루어졌다. 943명의 사용자들 개개인에게 가장 적합한 943개의 추천 시스템을 각각의 트레이닝 데이터셋과 DMLSVM 기법을 이용하여 구축한다. 이후 각 사용자별로 10개의 미지의 아이템에 대해서 해당 아이템을 추천할지 안할지를 결정한다. 10개의 아이템에 대한 rating 정보를 기반으로 추천시스템의 정확도를 평가한다. 각각의 제안하는 추천 시스템들의 테스트 데이터셋에 대한 목표값들은 표5에 도시하였다. 참고로 혼합 추천 시스템의 경우 평점 정보 3을 갖고 있는 아이템에 대한 분류 기준을 명확히 할 수 없기 때문에 혼합 추천 시스템은 모호한 데이터를 포함하지 않는 데이터셋에서만 평가의 검증이 가능하다.

표 5. 각 추천 시스템의 추천 목표값

Table 5. Target values of the propose RS.

	Target rating value for given RS				
	1	2	3	4	5
Positive RS	-1	-1	-1	1	1
Negative RS	-1	-1	1	1	1
Combined RS	-1	-1	N/A	1	1

제안하는 추천 시스템에서는 추천의 정확도를 향상시키기 위해서 3가지 종류의 커널 함수를 사용하였다. 3가지 함수는 쌍곡선 탄젠트 커널 함수, 다항식 커널 함수, 가우시안 커널 함수로 다음과 같이 표현 가능하다.

$$k_h(\mathbf{x}_i) = \tanh(\kappa \mathbf{x}_i^T \mathbf{x}_j + c) \quad (16)$$

$$k_p(\mathbf{x}_i) = (\mathbf{x}_i^T \mathbf{x}_j)^d$$

$$k_g(\mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

여기서  $k_h$ ,  $k_p$ ,  $k_g$ 는 각각 쌍곡선 탄젠트 커널 함수, 다항식 커널 함수, 가우시안 커널 함수를 나타낸다. 실험은 두가지 종류의 테스트 데이터셋을 가지고 하였다: 1) 위에서 설명한 각 사용자별 10개의 테스트 데이터셋 전부, 2) 모호한 데이터를 제외한 테스트 데이터셋. 여기서 모호한 데이터들은 평점 정보 3을 갖는 것들로 사용자의 호불호가 불명확한 데이터들을 의미한다. 실험 결과는 표 6과 같다. 실험 결과 어떤 커널을 사용하는지 여부에 관계없이 테스트 데이터셋 전부를

적용한 경우 적극적 추천 시스템이 소극적 추천 시스템보다 좋은 성능을 나타내는 반면 모호한 데이터들을 제외했을 경우에는 소극적 추천 시스템이 더 좋은 성능을 나타내는 것을 볼 수 있다. 이를 통해 유추해봤을 때 호불호가 분명한 아이템들의 경우 소극적 추천 시스템이 보다 사용자의 성향을 잘 예측하는 반면 모호한 아이템들의 경우 (본 실험에서는 평점을 3점 받은 영화들) 적극적 추천 시스템의 예측이 보다 정확하다고 볼 수 있겠다. 또한 모호한 데이터들이 제외된 경우 혼합 추천 시스템이 가장 좋은 성능을 보였다. 따라서 혼합 추천 시스템에서는 적극적 추천 시스템과 소극적 추천 시스템이 상호 보완적으로 decision making이 이루어진다고 할 수 있다. 참고로 모호한 데이터에 대한 혼합 추천 시스템의 결과는 표 7과 같다. 혼합 추천 시스템에서는 모호한 데이터를 대부분 추천하는 것을 볼 수 있다. 전체적인 실험결과로 미루어 보아 모호한 데이터가 없는 경우 가우시안 커널을 사용하는 혼합 추천 시스템을 사용하고 모호한 데이터가 포함된 경우는 쌍곡선 탄젠트 커널을 사용하는 적극적 추천 시스템을 사용하는 것이 안정적인 결과를 나타낼 것이다. 또한 모호한 데이터일지도 모르는 미지의 아이템이 제시되었을 경우 자신이 싫어하는 것에 대한 호불호가 명확한 사용자의 경우는 혼합 추천시스템을 사용하고 그렇지 않은 사용자의 경우는 적극적 추천시스템을 사용하면 사용자에게 가장 적합한 추천 결과를 얻을 것으로 보인다.

표 6. 제안된 추천 시스템의 실험 결과.  
Table 6. Simulation results of the proposed RS.

Method		Accuracy
Hyperbolic tangent kernel ( $\kappa=1, c=-1$ )	Positive RS (all data)	0.6450
	Negative RS (all data)	0.5892
	Positive RS (w/o ambiguous data)	0.7193
	Negative RS (w/o ambiguous data)	0.7915
	Combined RS (w/o ambiguous data)	0.9196
Polynomial kernel ( $d=1$ )	Positive RS (all data)	0.6034
	Negative RS (all data)	0.5756
	Positive RS (w/o ambiguous data)	0.6988
	Negative RS (w/o ambiguous data)	0.7748
Gaussian kernel ( $\sigma=1$ )	Positive RS (all data)	0.6355
	Negative RS (all data)	0.5880
	Positive RS (w/o ambiguous data)	0.7250
	Negative RS (w/o ambiguous data)	0.7931
	Combined RS (w/o ambiguous data)	0.9202

표 7. 모호한 데이터에 대한 혼합 추천 시스템 결과.  
Table 7. Results of combined RS for ambiguous data.

Kernel type	Number of ambiguous data	
	Recommend	Not recommend
Hyperbolic tangent	2187	237
Polynomial	2021	403
Gaussian ( $\sigma=1$ )	2218	206

마지막으로 제안하는 DMLSVM을 이용한 추천 시스템의 우수성을 입증하기 위해서 SVM [14], NN [15] 등의 결과와 비교하여 실험을 진행하였다. 실험 결과 제안하는 DMLSVM을 이용한 추천 시스템이 기존의 방법들보다 우수한 것을 표 8에서 볼 수 있다.

표 8. 제안된 추천 시스템과 기존 방법의 비교.  
Table 8. Comparison of previous and proposed methods.

Method		Accuracy
Proposed method (with Gaussian kernel)	Positive RS (all data)	0.6355
	Negative RS (all data)	0.5880
	Positive RS (w/o ambiguous data)	0.7250
	Negative RS (w/o ambiguous data)	0.7931
	Combined RS (w/o ambiguous data)	0.9202
Neural network	Positive RS (all data)	0.6172
	Negative RS (all data)	0.5725
	Positive RS (w/o ambiguous data)	0.6631
	Negative RS (w/o ambiguous data)	0.7706
SVM	Combined RS (w/o ambiguous data)	0.7878
	Positive RS (all data)	0.4599
	Negative RS (all data)	0.4186
	Positive RS (w/o ambiguous data)	0.4366
	Negative RS (w/o ambiguous data)	0.5634
Combined RS (w/o ambiguous data)	0.5764	

### 5. 결 론

본 논문에서는 커널 함수를 사용하는 새로운 방법의 추천 시스템을 제안하였다. 제안하는 방법은 유명한 기계 학습 모델 중 하나인 서포트 벡터 머신을 추천 시스템에 적용하였다. 제안하는 추천 시스템은 사용자에게 어떤 아이템이 적절한지를 추천하는 과정에서 해당 사용자의 과거의 평점 정보를 학습하여 이용하였다. 제안하는 방법에서 기계 학습 기법은 중대한 역할을 차지한다고 할 수 있다. 제안하는 방법의 기본 아이디어는 추천 문제를 이진 분류 문제로 공식화하고 이를 커널 함수를 이용하는 기계 학습 기법을 이용하여 푸는 것이

다. 기계 학습 기법중 DMLSVM을 이용하여 적극적 추천 시스템, 소극적 추천 시스템, 혼합 추천 시스템 등의 3가지 방법을 제안하였다. 제안하는 방법은 추천 시스템 연구분야에서 유명한 데이터베이스 중 하나인 MovieLens 데이터베이스에 적용하여 실험이 이루어졌다. 실험에서는 두 가지 종류의 테스트 데이터셋에 대하여 성능을 평가하였다: 1) 모든 데이터셋을 다 사용하는 경우, 2) 모호한 데이터를 제외한 경우. 실험결과 두 경우 모두 제안하는 추천 시스템이 우수한 성능을 보임을 알 수 있었다.

### 참 고 문 헌

[1] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *IEEE Internet Computing*, Jan./Feb., pp. 76-80, 2003.

[2] B.N. Miller, I. Albert, S.K. Lam, J.A. Konstan, and J. Riedl, "MovieLens unplugged: experiences with an occasionally connected recommender system," in *Proc. Int'l Conf. Intelligent User Interfaces*, pp. 263-266, 2003.

[3] D. Billsus, C.A. Brunk, C. Evans, B. Gladish, and M. Pazzani, "Adaptive interfaces for ubiquitous web access," *Comm. ACM*, vol. 45, no. 5, pp. 34-38, 2002.

[4] 김재광, 윤태복, 김동문, 이지형, "시간 가중치와 가변형 K-means 기법을 이용한 개인화된 음악 추천 시스템," *한국 지능시스템학회 논문지*, vol. 16, no. 2, pp. 504-510, 2009.

[5] 이세일, 이상용, "컨택스트 기반 협력적 필터링을 이용한 추천 시스템," *한국 지능시스템학회 논문지*, vol. 21, no. 2, pp. 224-229, 2011.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.

[7] N. Belkin and B. Croft, "Information filtering and information retrieval," *Comm. ACM*, vol. 35, no. 12, pp. 29-37, 1992.

[8] M. Pazzani and D. Billsus, "Learning and revising user profiles: the identification of interesting web sites," *Machine Learning*, vol. 27, pp. 313-331, 1997.

[9] M. Doumpos, C. Zopounidis, and V. Gollinopoulou, "Additive support vector machines for pattern classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 3, pp. 540-550, 2007.

[10] C.A. Waring and X. Liu, "Face detection using spectral histograms and SVMs," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 467-476, 2005.

[11] F. Camastra, "A SVM-based cursive character recognizer," *Pattern Recognition*, vol. 40, no. 12, pp. 3721-3727, 2007.

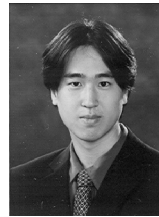
[12] J. P. Hwang, S. Park, E. Kim, and M. Park, "Dual margin approach on a lagrangian support vector machine," *International Journal of Computer Mathematics*, vol. 88, no. 4, pp. 695-708, 2011.

[13] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machines," *J. Mach. Learn. Res.*, vol. 1, pp. 161-177, 2001.

[14] V. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[15] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, 1999.

### 저 자 소 개



#### 이제현 (Jaehun Lee)

2005년 연세대학교 전기전자공학과 졸업(공학사)

2007년 연세대학교 전기전자공학과 석사과정 졸업(공학석사)

2007년~현재 연세대 전기전자공학과 박사과정

관심분야 : Computational intelligence, Localization and tracking in sensor network

E-mail : aznable17@yonsei.ac.kr



#### 황재필 (Jae Pil Hwang)

2003년 연세대학교 전기전자공학부 졸업(공학사)

2005년 연세대학교 전기전자공학과 석사과정 졸업(공학석사)

2011년 연세대학교 전기전자공학과 박사과정 졸업(공학박사)

2011년 3월~현재 현대자동차그룹 책임연구원

관심분야 : 지능형 차량, 로봇 비전

E-mail : purnnara@yonsei.ac.kr



#### 김은태 (Euntai Kim)

1992년 연세대학교 전자공학과 졸업(공학사)

1994년 연세대학교 전자공학과 석사과정 졸업(공학석사)

1999년 연세대학교 전자공학과 박사과정 졸업(공학박사)

1999년 3월~2002년 2월 국립한경대학교 제어계측공학과 조교수

2002년 3월~현재 연세대학교 전기전자공학부 부교수

2003년 University of Alberta, visiting researcher

1998년~현재 IEEE TFS, IEEE SMC, IEEE CAS, FSS 등에서 심의위원 활동 중

2003년 대한 전자공학회 해동상 수상

관심분야 : Computational intelligence, 지능형 로봇

Phone : +82-2-2123-2863

E-mail : etkim@yonsei.ac.kr