앙상블 SVM을 이용한 동적 웹 정보 예측 시스템

박 창 희[†] · 윤 경 배[††]

요 약

기존의 웹 정보 예측 시스템은 예측에 필요한 정보를 얻기 위하여 사용자 프로파일과 사용자로부터의 명시적 피드백 정보를 필요로 하는 단점이 존재한다. 본 논문에서는 이러한 단점을 극복하고자 웹 사이트에 접속한 고객의 행동을 나타내는 클릭 스트림 데이터와 이를 기반으로 한 사용자의 암시적 피드백 정보를 이용하여 각 사용자가 가장 필요로 하는 웹 정보를 예측한다. 이를 이용하여 관련 정보를 제공할 수 있는 앙상블 SVM을 이용한 동적 웹 정보 예측 시스템을 설계하고 구현하며, 기존의 웹 정보 예측 시스템과 성능 비교를 수행한 결과, 제안된 방법의 우수함이 입증되었다.

Dynamic Web Information Predictive System Using Ensemble Support Vector Machine

Changhee Park[†] · Kyungbae Yoon[††]

ABSTRACT

Web Information Predictive Systems have the restriction such as they need users' profiles and visible feedback information for obtaining the necessary information. For overcoming this restrict, this study designed and implemented Dynamic Web Information Predictive System using Ensemble Support Vector Machine to be able to predict the web information and provide the relevant information every user needs most by click stream data and user feedback information, which have some clues based on the data. The result of performance test using Dynamic Web Information Predictive System using Ensemble Support Vector Machine against the existing Web Information Predictive System has proved that this study's method is an excellence solution.

키워드 : 앙상블(Ensemble), 예측(Predictive), SVM(Support Vector Machne), 마이닝(Mining), 선택(Voting), 분류(Classification)

1. 서 론

1990년대 이후 WWW의 확장과 인터넷의 대중화는 기하급수적인 정보량의 증가를 가져왔고 네티즌들은 전세계에 분산되어 있는 정보들을 매우 적은 비용으로 얻을 수 있게 되었다. 뿐만 아니라 인터넷의 발전은 전반적인 사회 구조의 변화까지 일으켰으며, 이러한 발전은 인류의 생활 패턴을 바꾸었고, 전자상거래와 같은 새로운 산업 구조가 도래하였으며, 인터넷을 떠나서는 살아갈 수 없을 정도로 인간 활동의 상당부분을 인터넷에 의존하게 되었다. 그러나 Dizzy Web이라 부를 수 있을 정도로 무질서한 현 인터넷 환경에서 특정 정보를 찾고자 무작정 웹을 검색한다는 것은 매우 비효율적인 행위가 아닐 수 없다. 따라서 정보의 구조화를 통해 웹 사이트의 재정비를 꾀하고, 현재 인터넷 웹 사이트의 구조적 비효율성을 제거하여 사용자로 하여금 좀더 쉽고 경제적으로 정보를 얻을 수 있는 방안 제시에 관한 연구가 절실하다[1, 3].

이를 위해 가장 필요한 연구 분야는 사용자 모델링(user modeling)을 들 수 있으며, 이는 사이트를 찾아온 사용자가 어떤 부류에 속하는지, 이용하는 패턴 및 전반적인 성향은 어떠한지를 구체화하여 이를 시스템에서 이용할 수 있는 형태로 표현하는 분야로써, 사용자의 성향을 나타낼 수 있는 많은 정보들, 즉 사용자 프로파일, 웹 로그, 통계학적 정보 등의 방대하고 기초적인 데이터들로부터 유용한 정보들을 추출하여 시스템에서 이용 가능한 형태로 재구축하는 작업들을 대상으로 연구가 진행되고 있다. 그 중 웹 마이닝은 웹에서 필요한 정보를 발견하고 추출하는 과정이라 정의할 수 있으며, 내용(contents), 구조(structure), 사용(usage) 마이닝으로 분류할 수 있다. 이 중 사용 마이닝은 사용자가 사이트에 접속한 로그 파일을 가지고 결과를 분석하는 것으로 개인화, 시스템 성능향상, 사이트 수정 및 사용자 특성화 등에 주로 활용되고 있다[4].

그러나 기존의 웹마이닝 방법은 한 개의 웹 사이트에 접

† 준 회 원 : 연세대학교 대학원 전자공학
†† 정 회 원 : 김포대학 컴퓨터계열 교수
논문접수 : 2003년 12월 1일, 심사완료 : 2004년 5월 24일

속한 사용자가 해당 사이트의 전체 페이지를 모두 볼 수는 없기 때문에 한번의 접속으로 사용자가 방문한 웹 페이지보다도 머무르지 않고 그냥 지나친 페이지의 수가 훨씬 많기 때문에 웹 데이터에 대한 희소성(sparsity) 문제가 발생한다[10]. 따라서 본 논문에서는 이러한 희소성있는 웹 로그 데이터를 기반으로 사용자로부터 얻은 연속성 피드백 정보를 이용하며, 최근 그 속도와 높은 정확성으로 패턴인식 연구 분야에서 다각도로 연구되고 있는 Support Vector Machine (SVM)에 우수 모델 선택(voting) 기법인 앙상블(Ensemble) 모형을 결합한 앙상블 SVM을 설계하여 사용함으로서 웹 로그 데이터의 희소성 문제를 해결한 앙상블 SVM을 이용한 동적 웹 정보예측 시스템을 제안하고 실험을 통하여 증명한다.

2. SVM

SVM은 1970년대 후반 Vapnik이 주어진 데이터 (x_1, y_1), (x_2, y_2), …, (x_n, y_n)을 이분법적으로 가장 잘 나눌 수 있는 선형평면을 구하는 방법으로 제안하였다[12, 13].

임의의 데이터 $\{(X_i, Y_i),\ i = 1, \cdots, N\}$가 주어졌을 때, X_i는 두 클래스 중 하나에 속하며, $Y_i \in \{-1, 1\}$는 해당 클래스를 표시하는 역할을 한다. SVM은 각 클래스를 구분하는 최적의 분리 경계면을 구하기 위해 분리 경계면과 분리 경계면에 가장 인접한 점과의 거리를 최대화한다. 최적의 선형 분리 경계면을 $f(x) = W^T X + b$로 놓으면, support vector와 $f(x)$의 거리를 $1/\|w\|$로 나타낼 수 있다. SVM은 $\|w\|^2$를 최소화하여 분리 간격을 최대화하도록 하여 최적 분리면을 찾아내며, 다음과 같은 블록 최적화 문제가 된다.

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 \\ &y_i(w^T x_i + b) \geq 1 \quad \text{for } i = 1, \cdots, N \end{aligned} \tag{1}$$

이 문제를 라그랑제(Lagrange) 배수로써 쌍대화(Dual Problem) 시키면 식 (2)과 같이 된다.

$$\begin{aligned} &\Theta(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \langle x_i, x_j \rangle \\ &\text{s.t. } \alpha_i \geq 0,\ i = 1, \cdots, N \text{ and } \sum_{i=1}^{N} \alpha_i d_i = 0 \end{aligned} \tag{2}$$

선형 분리경계면으로 완전히 구분할 수 없는 서로 겹쳐져 있는 패턴의 경우에는 슬랙변수 ξ를 적용하여, 식 (1)로부터 다음과 같이 나타낼 수 있다.

$$\begin{aligned} &\text{minmize } \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ &\text{s.t. } y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \cdots, N \end{aligned} \tag{3}$$

위 식 (3)의 $y_i(w^T x_i + b) \geq 1 - \xi$에서 $\xi_i = 0 (\forall i)$이면 모든 패턴을 완전하게 분리할 수 있다는 것을 의미한다. 그러나 대부분의 패턴은 선형적으로 분리가 가능하지 않다. 따라서 비선형 패턴을 분리하기 위하여 비선형 패턴의 입력 공간을 선형 패턴의 특징 공간으로 전환한다.

즉, 커널함수 $K(X_i, X_j) = \phi(X_i)\,\phi(X_j)$를 정의하면 비선형 패턴을 분리하기 위한 모델은 식 (1)~식 (3)으로부터 아래와 같이 표현된다.

$$\begin{aligned} &\theta(\alpha) = \sum_i \alpha_i^N - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(x_i, x_j) \\ &\text{s.t. } \sum_i^N \alpha_i d_i = 0,\ 0 \leq \alpha_i \leq C,\ \forall i. \end{aligned} \tag{4}$$

여기서 C는 식 (2)에서의 Penalty parameter이다. 위의 모델에서 라그랑제 배수 i를 구하면 특징 공간에서 가장 평평한 함수인 아래 식 (5)을 구할 수 있다.

$$\begin{aligned} f(x) &= \text{sign}(\langle w, \phi(x) \rangle + b) \\ &= \text{sign}\Big(\sum_{i=1}^{N} \alpha_i d_i K(x_i, x) + b\Big) \end{aligned} \tag{5}$$

이때 커널함수 K는 여러 함수 중 선택될 수 있다.

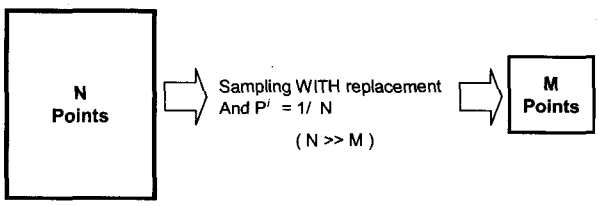
SVM은 빠른 학습 시간 때문에 동적인 예측 모형에서 사용되지만 예측의 정확성에 있어 기존의 웹 예측 시스템에서 주로 사용하고 있는 피어슨의 상관 계수 알고리즘에 비해 향상된 성능을 보이지는 못한다.

본 논문에서는 이러한 문제의 해결 방안으로 7개의 커널 함수들을 사용하여 주어진 학습 데이터를 가장 잘 모형화하는 한 개의 커널함수를 찾아내는 앙상블 기법을 사용한다.

3. 앙상블 SVM

기존의 SVM 모형은 한 개의 커널 함수를 사용하지만 웹 데이터 분석을 위한 앙상블 SVM 모형은 다수의 커널 함수를 사용하여 가장 좋은 성능을 나타내는 함수를 voting하여 적용하는 앙상블 기법을 결합하여 최적의 모형을 구하게 된다. 기존의 Voting하는 방법으로는 bagging, boosting, randomization, stacking, dagging 등이 있다[5, 7, 9, 11]. 본 논문에서는 서로 다른 커널 함수를 갖는 SVM 학습 모형의 군집들 중에서 학습 과정을 통해 가장 작은 평균 제곱 오차(MSE : Mean Squared Error)를 갖는 모형을 결정하는 방법을 수행한다. 이 방법에 의하여 기존의 SVM 보다 더 정확한 예측력을 갖는 모형을 얻게 된다. 그러나 앙상블 모형을 수행하기 위해서는 앙상블을 사용하지 않는 모형에 비해서 더 많은 학습을 하게 되어 학습 시간에 대한 비용이 증가하게 된다. 이러한 앙상블 모형의 문제점을 해결하기 위해 앙상블 SVM은 (그림 1)의 구조를 갖는 붓스트래핑(Bootstrapping) 기법을 사용한다. 이 방법은 모형의 학습을 위하여 전체 데이터를 사용하지 않고 반복이 있는(with replacement)

임의 추출 방법인 재표본(resampling) 기법을 사용한다. 이때 데이터의 크기가 N인 전체 데이터의 각 개체(point) $Point_N = \{(x_i, y_i) \mid i = 1, 2, \cdots, N\}$은 모든 같은 확률(1/N)로서 추출되어, $Point_M = m = 1, 2, \cdots, M$으로 M개의 크기 표본으로 이루어지고 이 표본을 통해 앙상블 학습이 이루어진다. 이때 전체 데이터의 크기 N에 비해서 앙상블 학습에 사용되는 표본의 크기인 M은 매우 작도록 한다.



(그림 1) 붓스트래핑

앙상블 SVM의 알고리즘은 (알고리즘 1)과 (알고리즘 2)의 두 단계로 구성된다. (알고리즘 1)은 앙상블 SVM의 초기화 단계로서, X는 입력 변수, Y는 목표 변수, ker은 커널 함수, C는 상한(upper bound), *loss*는 손실 함수, e는 insensitivity, *nsv*는 support vector의 개수, *beta*는 라그랑지곱 차분, 그리고 *bias*는 바이어스 항을 나타낸다. 여기서 커널 함수가 다양한 형태로 바뀌면서 학습이 이루어지는데, 처음 결정하는 모수들이 제대로 결정되었는지를 조사하는 프로세스를 거치게 된다. (알고리즘 2)는 (알고리즘 1)에 의하여 필요한 모수가 결정된 후, 최적의 커널 함수를 결정하는 앙상블 단계이며, 이 과정을 통하여 SVM의 정확도에 대한 성능 향상이 이루어진다.

```
Algorithm : Initialize_앙상블 SVM (Parameter[j])
// 앙상블 SVM의 초기화
Input : f [nsv, beta, bias] = svr (X, Y, ker, C, loss, e)
Output : H = kernel(ker, X(i), X(j))
// 매개변수의 초기값 조사
if (arguments < 3 | arguments > 6)
// 매개 변수의 정확한 개수 조사
else n = size (X, 1)
        // 입력 데이터 확인
        if (arguments < 5) loss = eInsensitive
        // 초기 손실함수를 ε -Insensitive로 결정
        if (arguments < 4) C = Inf
        // 학습 모형의 모수 추정의 상한 결정
        if( arguments < 3) ker = linear
        // 초기 커널 함수를 linear로 결정
end
// 커널 함수의 결정
Set H = zeros (n,n)          // 커널 함수 배열의 초기화
for (i =1 ; i<=n ; i++)
        for (j =1 ; j<=n ; j++)
                H(i, j) = kernel (ker, X(i), X(j)) // 커널함수의 적용
end
```

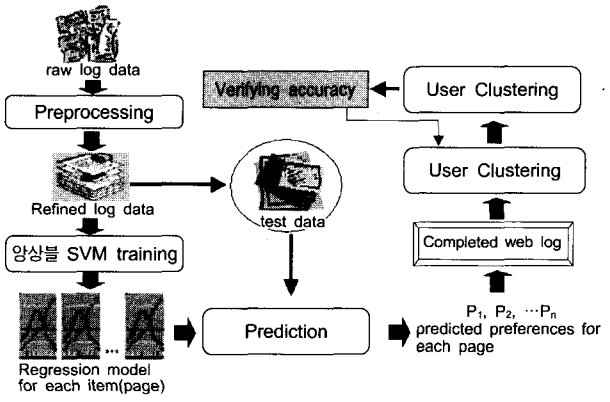
(알고리즘 1) 앙상블 SVM 초기화 단계

```
Algorithm : Voting_Kernel(Select_kernel[k])
// 다수의 커널 함수를 사용하여 최적의 모수를 결정
Choose optimal kernel K* such that min||w||
// 1-scatter smoothing, 2-bin, 3-running mean,
// 4-kernel smoother, 5-equivalent kernel,
// 6-regression spline, 7-cubic smoothing spline

// 붓스트랩 샘플링
for (i = 1 ; i =< 13109 ; i++)
        if random_number < 0.1
                re_sampling
// 붓스트래핑의 재 표본 기법 적용
end

// 최적 커널 함수의 결정
for (k = 1 ; k=<7 ; k++)
        MSE[k]=risk(Select_kernel[k])
// 위험함수 값의 최소 제곱 오차 계산
        if MSE[k] = min ;
         voting = MSE[k]
// 최소제곱오차(MSE)값이 가장 작은 커널함수를 선택
end
```

(알고리즘 2) 최적 커널 함수 결정을 위한 앙상블 단계

4. 앙상블 SVM을 이용한 동적 웹 정보 예측 시스템 설계

앙상블 SVM을 이용한 동적 웹 정보 예측 시스템은 웹 서버에 접속하는 사용자에게 가장 필요한 정보를 예측하여 실시간으로 해당 정보를 추천해 주는 웹 정보 추천 시스템으로, 웹 서버에 저장되어 있는 웹 로그 데이터로부터 전처리 과정을 거쳐, 제안하는 앙상블 SVM 알고리즘을 사용하여 웹 정보에 대한 예측 모형을 만들어 이 모형에 의해 새로운 사용자에게 적합한 정보를 추천하게 된다.



(그림 2) 동적 웹 추천시스템 설계

(그림 2)는 설계 절차로서 여러 단계의 절차를 거쳐 수행되며, 각각의 단계는 다음과 같다. 첫째 단계로 웹 서버의 로그 데이터로부터 사용자들의 클릭 스트림 정보를 얻는다. 일반적으로 웹 로그 데이터는 로그를 생성한 서버의 종류와 웹 사이트의 성격, 그리고 사이트를 제공하는 주체의 성격에

따라 다양한 형태를 나타낸다. 따라서 초기의 로그 데이터는 전처리 과정을 통하여 모델을 구축하는 데 있어 필요한 정보만을 적절한 형태로 추출하게 된다. 이렇게 정제된 로그 정보는 사용자가 방문한 페이지와 페이지에 머문 시간을 모아서 하나의 인스턴스로 이용하고, 여기서 적절한 수의 인스턴스를 다시 추출하여 학습 데이터로 사용함으로서 사용자에 대한 웹 페이지 선호도에 대한 예측 모형을 만든다. 둘째 단계는 테스트 데이터를 이용하여 얻은 정보를 전체 웹 페이지의 앙상블 SVM 모델에 적용시켜 각 페이지에 대한 사용자의 선호도를 예측해 낸다. 이때의 테스트 데이터는 정제된 로그파일에서 학습 데이터에 포함되지 않은 데이터들 중에 무작위로 추출하여 성능 평가에 사용한다. 계속해서 앞의 각 과정들을 통해 얻은 각 페이지들에 대한 예측 값에 각 사용자의 평균 관심도와 각 페이지의 평균 관심도를 고려하여 식 (6)과 같이 선호도확률을 계산한다.

$$P(\text{페이지선호도}\mid\text{입력페이지}) = \frac{P(\text{사용자의페이지선호도})\times P(\text{입력페이지로인한비율})}{\sum_{\text{페이지}=0}^{k} P(\text{사용자의페이지선호도})\times P(\text{입력페이지로인한비율})} \quad (6)$$

이 과정을 통해 사용자가 방문하지 않은 페이지도 모두 예측되어 결측치가 없는 완전한 데이터 구조를 이루게 된다.

이러한 자료의 특징은 데이터가 매우 희소하다는 것이다. 즉 한 고객이 해당 웹 사이트에 접속하여 보게 되는 페이지의 수는 매우 적기 때문에 접속을 하지 않은 페이지에 대한 값은 모두 측정이 안된 결측치로 처리된다. 일반적인 데이터 마이닝 알고리즘으로는 이러한 데이터에 대한 모형을 만들기가 어렵다. 이러한 문제점을 해결하기 위하여 본 논문에서 제안하는 앙상블 SVM 모형을 제안하였고, 이 논문의 실험 데이터에 적용하여 희소한 웹 로그의 클릭 스트림 데이터를 결측치가 없는 완전한 데이터 구조로 만들었다.

또한 본 논문에 대한 측도(metric)는 두 가지를 사용하였다. 첫 번째 측도는 앙상블 SVM 모형의 유효성을 평가하기 위하여 이 모형의 출력결과를 주어진 값과 비교하여 그 모델의 정확성을 살피기 위해 MSE를 이용하였고, 두 번째로 사용된 측도는 모형을 이용한 추천 시스템의 성능을 측정하기 위하여 순위 비율(ranking rate)을 사용하였다. 순위 비율은 사용자의 만족 여부를 나타내는 측도로서 사용자 a의 페이지 k에 대한 만족 정도는 식 (7)과 같이 표현된다.

$$R_{ak} = \frac{P_{ak}}{r_a} \cdot \frac{P_{ak}}{r_k} \quad (7)$$

여기서 r_a는 a의 순위, r_k는 k의 순위, P_{ak}는 사용자 a가 페이지 k에 대한 선호도를 나타낸다. R_{ak}의 값이 1이상일 경우에는 만족함을 나타낸다. 반면에 이 값이 1미만일 경우에는 그렇지 못함을 의미한다. 실험에서는 이를 이용하여 각 학습 데이터에 포함되지 않은 모든 페이지에 대하여 예측을 실시하고 순위 비율을 계산하여 선호하는 페이지의 순위를 부여하게 된다.

5. 실험 및 평가

본 논문에서는 웹 로그 데이터로부터 웹 정보 추천에 적용되는 앙상블 SVM을 이용한 동적 웹 정보 예측 시스템의 성능을 평가하기 위해 실제 인터넷 쇼핑몰의 웹 로그 데이터를 사용하였다. 이 논문의 실험에서 사용한 데이터는 2000년도 KDD Cup 대회에 참가한 참가자들에게 문제로 주어졌던 웹 로그의 클릭 스트림 데이터이며, 이 데이터는 인터넷 쇼핑몰 Gazelle.com의 2개월간의 클릭 스트림 만을 모아 놓은 1.2GB의 텍스트 데이터이다[6]. 실험은 Pentium-IV 2GHz CPU와 1G RAM, Windows XP 환경에서 마이크로 소프트 Visual C++ 6.0으로 구현하였다.

실험에 사용된 데이터의 전처리는 유효 사용자 추출, 특성 애트리뷰트 데이터 추출, 시간 데이터 변환의 3가지 과정을 거친다. 우선 사용자의 구분은 쿠키(cookie) ID를 이용하여 부여하였다. 페이지 방문시간의 계산은 페이지 요청 처리(page request processing)시간부터 다음 클릭 스트림이 발생할 때까지의 시간 간격으로 계산하였고, 만약 다음 클릭 스트림이 해당 사용자가 아닐 경우에는 그 때까지 클릭 스트림에 대하여 계산된 방문 시간의 평균을 부여하였다. 한편 세션(session)이 다를 경우에는 같은 쿠키 ID를 지닌 사용자일지라도 별도의 사용자로 가정하고 시간을 계산하였다. 그리고 상한 임계값은 1,000초로 설정하였다.

본 논문에서 정확성 성능 평가에 대한 MSE는 전체 데이터에 대한 것과 상위 50%에 대한 것으로 나누어 계산하였다.

〈표 1〉 앙상블 SVM을 이용한 웹 예측 성능 평가

(단위 : 초)

구 분	Pearson	SVM	앙상블 SVM
MSE(전체)	1.37	1.29	0.89
MSE(상위 50%)	1.01	0.97	0.64

<표 1>은 전체 페이지 모델에 대하여 페이지 별로 각각 생성된 테스트 데이터에 대한 결과들의 평균 MSE값이다. 이 표에서 보듯이 전반적인 정확도에 있어 Pearson의 상관계수 알고리즘과 SVM은 큰 차이를 보이지 않고 서로 비슷함을 알 수 있다. 하지만 앙상블 SVM에 의한 모형의 정확도는 앞의 두 알고리즘에 비해 훨씬 정확함을 알 수 있다. 이렇게 기존의 두 개의 모형보다 제안하는 앙상블 SVM이 정확하게 나올 수 있는 이유는 이 모형이 데이터를 학습하는 동안 최적 커널 함수를 선택하는 앙상블 구조를 가지고 있기 때문이다.

<표 2>는 데이터의 크기가 증가함에 따라 학습 시간의

차이를 비교하였다.

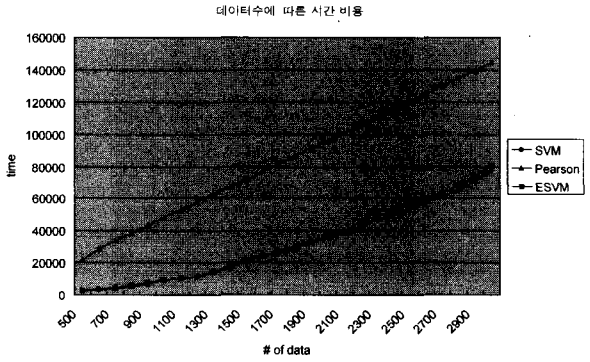
〈표 2〉 데이터 량에 따른 학습 시간

(단위 : 10^{-3}초)

데이터 크기	Pearson	SVM	앙상블 SVM
500	21,873	2,941	2,955
1,000	48,121	8,890	8,954
1,500	72,181	20,631	22,009
2,000	96,241	35,170	36,542
2,500	120,302	54,391	57,325
3,000	144,362	77,967	81,024

현재 매우 빠른 학습 시간이 가장 큰 장점 중 하나로 갖는 SVM 모형은 데이터의 크기가 증가함에도 빠른 학습 시간을 보이고 있다. 특히 데이터의 수가 증가하면 할수록 피어슨의 상관 계수 알고리즘에 비해 훨씬 빠른 학습 시간을 보여 주고 있다. 앙상블 SVM도 SVM에 비해 결코 떨어지지 않는 시간 성능을 보이고 있다.

(그림 3)은 500개의 데이터 크기부터 3,000개 데이터까지 100개씩 데이터를 증가시켜서 학습한 학습 시간의 결과이다.



(그림 3) 3가지 모형간의 학습 시간 비교(단위 : 10^{-3}초)

(그림 3)의 데이터 수에 따른 학습 수행 후 예측 시간에 있어 피어슨의 상관 계수 알고리즘은 앙상블 SVM방식에 비해 매우 느림을 실험을 통해 알 수 있었다. (그림 3)에서 보듯 SVM은 피어슨의 상관 계수 알고리즘에 비해 탁월한 시간 감소효과를 보인다. 하지만 SVM과 앙상블 SVM은 별 차이를 보이지 않고 있다. 데이터의 크기가 1,000이하일 경우에는 정확도가 매우 떨어져 의미가 없긴 하나, 데이터 크기에 따른 알고리즘들의 시간비용을 측정하기 위하여 이 부분까지 포함하여 제시하였다. 특히 앙상블 SVM 모델이 제시하는 시간비용은 학습 시간을 포함한 것이다. 실제 학습된 모델을 이용하여 선호도를 예측할 때는 실시간 내(평균 0.1초 이내)에 동작하였다.

5.1 웹 페이지 추천 시스템의 성능 평가

이 실험은 앙상블 SVM 기반 사용자 모델링을 이용한 추천 시스템의 성능을 평가하는 실험이다. 4,457명의 테스트 데이터의 10% 정도를 이용하기 위하여 각 사용자 별로 1회의 방문을 통한 웹 페이지의 수를 조사해 보니 14페이지 이상 방문한 사용자가 498명이었다. 따라서 이 절의 실험은 전체 269개의 웹 페이지 중에서 14페이지 이상 방문한 498명을 대상으로 7개의 페이지는 이미 방문한 페이지로 설정하고, 나머지 7개의 페이지에 대하여 예측한 선호도를 이용해 순위 비율을 구하고 순위 비율의 순서대로 예측 페이지의 HIGH 선호 아이템과 LOW 선호 아이템을 선택한 후, 실제 데이터에서 보이는 선호도와 비교하였다. 그 결과는 <표 3>과 같이 나타났다.

〈표 3〉 순위 비율을 이용한 시스템의 성능 평가

구 분	Pearson	SVM	앙상블 SVM
Pr(high/high)	0.34	0.32	0.41
Pr(low/low)	0.29	0.30	0.43
Pr(high/low)	0.17	0.18	0.10
Pr(low/high)	0.11	0.12	0.08

앙상블 SVM은 피어슨의 상관 계수 알고리즘과 SVM에 비해 훨씬 정확한 결과를 보이고 있다. 웹 페이지의 순위 비율이 높은 웹 페이지에 대하여 같은 high의 웹 페이지가 나올 확률이 다른 모형에 비해 훨씬 높았다. 순위 비율이 낮은 웹 페이지에 대한 낮은 것의 예측 확률도 마찬가지로 2개의 비교되는 다른 모형에 비해 높은 확률값을 보여주고 있다.

6. 결 론

본 논문에서 제안한 앙상블 SVM을 이용한 동적 웹 정보 예측 시스템은 웹 로그 데이터의 희소성 문제를 해결하고 웹 정보의 추천을 위하여 앙상블 SVM 기법을 적용한 시스템이다. 이는 사용자별로 적합한 웹 정보를 추천하는 일련의 과정이 실시간으로 이루어질 수 있도록 설계되었으며, 이렇게 설계된 시스템은 기존의 추천 시스템보다 정확도, 예측을 위한 학습 시간에서 더욱 우수함이 실험을 통하여 입증되었다.

그러나 제안한 시스템이 기존의 추천 시스템에 비해 우수한 성능을 보이고 있지만 예측의 정확성 문제점을 여전히 내포하고 있다. 따라서 향후에는 이러한 제약조건을 해결하기 위해 다른 데이터 마이닝 기법의 결합을 통하여 다양한 측면에서의 희소성 데이터를 분석하고 추천할 수 있는 새로운 알고리즘의 개발이 필요하며, 클릭 스트림 이외의 다양한 웹 로그 데이터를 이용하여 본 논문에서 제안하는 시스템을 적용함으로서 동적 웹 예측 기법의 기능을 지속적으로 확장할 수 있도록 연구를 수행해 나가야 할 것이다.

참 고 문 헌

[1] 윤경배, 최준혁, 왕창종, "하이브리드 SOM을 이용한 효율적

인 지식베이스 관리", 정보처리학회논문지B, 제9-B권 제5호, pp.46-53, 2002.

[2] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers," IEEE Transaction on signal processing, Vol.45, No.11, pp.2758-2765, 1997.

[3] D. Lewis, "Evaluating text caterorization," *Proceedings of Speech and Natural Language Workshop*, pp.122-132, 1991.

[4] E. Frank, "Data Mining," Morgan Kaufmann Publisher, 2000.

[5] H. C. Kim, S. Pang, H. M. Je, D. J. Kim and S. Y. Bang, "Constructing support vector machine ensemble," Pattern Recognition, 36, pp.2757-2767, 2003.

[6] "http://www.ecn.purdue.edu/KDDCUP/," 2003.

[7] K. Ting and I. Witten, "Stacking bagged and dagged models," Proceedings of the 14th International Conference on Machine Learning, Sapporo, Japan, 1997.

[8] M. Pontil and A. Verri, "Properties of Support Vector Machines," A. I. Memo No.1612, CBCL paper No.152. Massachusetts Institute of Technology, Cambridge, 1997.

[9] R. Maclin and D. Opitz, "An empirical evaluation of bagging and boosting," Proceedings of the fourth National Conference on Artificial Intelligence, Austin, Texas, 1997.

[10] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," VLDB, pp.407-419, 1995.

[11] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees : bagging, boosting, and randomization," Mach. Learn, 26, pp.1-22, 1998.

[12] V. N. Vapnik and C. Cortes, "Support vector networks," *Machine Learning*, 20, pp.273-297, 1995.

[13] V. N. Vapnik, "Statistical Learning Theory," John Wiley and Sons, 1998.

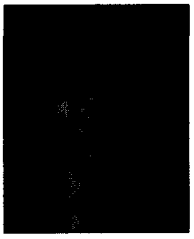
박 창 희

e-mail : chpark92@yonsei.ac.kr

2000년 한양대학교 전자과(학사)

2003년~현재 연세대학교 공학대학원 전자공학 석사과정

관심분야 : 생체(지문, 홍채, 음성)인식, 데이터마이닝, 인공지능, 영상처리 등



윤 경 배

e-mail : kbyoon@kimpo.ac.kr

1986년 인하대학교 수학과(학사)

1994년 인하대학교 대학원 정보공학과 (공학석사)

1998년 서강대학교 경제대학원 정보기술 경제학(경제학석사)

2003년 인하대학교 대학원 전자계산공학과(공학박사)

1986년~1987년 대우자동차(주) MIS 근무

1988년~1991년 LG-EDS(주) 기술연구소 근무

1992년~1997년 동부정보기술(주) 정보기술연구소 근무

1998년~현재 김포대학 컴퓨터계열 교수

관심분야 : 지식기반 데이터베이스, 데이터마이닝, CRM, 지문 및 음성 인식, 인공지능 등