The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 197-211

Collaborative CRM using Statistical Learning Theory and Bayesian Fuzzy Clustering

Sung-Hae Jun[1)]

Abstract

According to the increase of internet application, the marketing process as well as the research and survey, the education process, and administration of government are very depended on web bases. All kinds of goods and sales which are traded on the internet shopping malls are extremely increased. So, the necessity of automatically intelligent information system is shown, this system manages web site connected users for effective marketing. For the recommendation system which can offer a fit information from numerous web contents to user, we propose an automatic recommendation system which furnish necessary information to connected web user using statistical learning theory and bayesian fuzzy clustering. This system is called collaborative CRM in this paper. The performance of proposed system is compared with the other methods using real data of the existent shopping mall site. This paper shows that the predictive accuracy of the proposed system is improved by comparison with others.

Keywords: Statistical Learning Theory, Bayesian Fuzzy Clustering, Collaborative CRM.

1. 서론

웹 환경의 지속적인 확장은 정보화가 필요한 데이터의 크기를 빠르게 증가시켜 왔다. 인터넷을 이용하여 전 세계의 대부분의 정보를 자신의 현재 위치에서 매우 적은 비용으로 획득할 수 있게 되었을 뿐만 아니라 인터넷의 발전으로 인하여 전반적인 사회 구조의 변화를 가져와 사람들의 생활 패턴까지 바꾸었고 심지어 인터넷 중독증과 같은 악 영향까지도 가져왔지만 전자 상거래와 같은 새로운 산업 구조를 개척하는 등 장점들도 보이고 있다. 현재 많은 기업들이 인터넷 환경에서 자사의 사이트에 접속하는 사용자들에 적절한 웹 컨텐츠를 추천해 주는 시스템을 운영하고 있지만 사용자들의 다양한 정보화 요구를 만족시키는 시스템은 많지 않다. Forrester(2002)는 인터넷 사용자들이 실제적으로 자신들의 필요 정보를 찾아보는 시간은 전체 인터넷 사용시간의 42%에 지나지 않으며, 그 외의 시간은 전혀 불필요한 곳에서 정보를 찾기 위해 낭비하고 있다고 했다. 또한 전체 인터넷 사이트의 51%는 웹 페이지의 내용 구성이 이해하기 어렵게 되어 있으며, 결국 전체의 90% 이상이 적절치 못한 구조를 지닌다고 하였다. 디지 웹(dizzy web)이라 부를 수 있을

1) Full-time Instructor, Dept. of Statistics, Chongju University
E-mail : shjun@cju.ac.kr

정도로 무질서한 현재의 인터넷 환경에서 특정 정보를 찾기 위하여 무작정 웹 서핑을 한다는 것은 매우 비효율적인 작업이며, 따라서 정보 구조화를 통해 웹 사이트의 재정비를 꾀하고, 인터넷의 구조적 비효율성을 제거하여 사용자로 하여금 좀더 쉽고 경제적으로 정보를 얻을 수 있도록 하는 방안에 대한 연구가 Basu, et al.(1998)와 Cooley, et al.(1997) 등에 의해 이루어지고 있다. Fisher, et al.(2000) 등에 의해 연구되어지고 있는 개인화 웹(personalized web)은 앞의 문제를 해결하고자 하는 연구 분야중 하나로서, 웹 사이트에 접속하는 모든 사용자들에게 획일적인 정보를 제공하는 것이 아니라, 각종 정보들로부터 사용자의 성향을 파악하고, 이에 맞추어 사이트를 적응, 혹은 변화시켜 서비스를 제공한다. 즉, 개인화 웹의 연구는 해당 사이트로부터 좀더 쉽고, 빠르며, 효과적으로 사용자에게 적절한 정보를 제공하고자 하는 것이며 이를 통하여 궁극적으로 웹 사이트를 방문한 사용자로부터 보다 큰 만족도를 얻고자 하는 것이다. 웹 개인화는 전자 상거래와 같은 기업 활동에 있어서 더욱 강조되고 있는데, 대표적인 예를 아마존닷컴(amazon.com)에서 찾아볼 수 있다. 아마존의 창립자 Bezos는 "만약 우리의 웹 사이트에 1000만 명의 방문객이 있다면, 우리는 이들을 위해 1000만 개의 웹 사이트들을 만들겠다."라고 하였다(Forrester, 2002). 이는 웹의 기본적인 역할을 정보의 전달로 보았을 때, 전자 상거래를 위한 사이트는 물론이고, 기타 모든 사이트들에 대해서도 해당되는 서비스 제공 전략이라 할 수 있다. 웹 개인화를 위한 효과적인 전략의 하나로서 사용자 모델링(user modeling)이 있다. 사용자 모델링은 자신의 웹 사이트를 찾아온 사용자가 어떤 부류에 속하는지, 이용하는 패턴 및 전반적인 성향은 어떤지를 구체화하여 이를 시스템에서 이용할 수 있는 형태로 모형화 한다. 이러한 모형화 전략은 크게 웹 마이닝(web mining)의 범위에 속한다. 웹 마이닝은 사용자의 성향을 나타낼 수 있는 많은 정보들, 사용자 프로파일, 웹 로그, 통계학적 정보 등의 방대하고 기초적인 데이터들로부터 유용한 정보들을 추출하여 시스템에서 이용 가능한 형태로 재구축하는 작업들을 포함하고 있다. 사용자 모델링을 이용하여 협업 CRM(collaborative Customer Relationship Management)을 위한 추천 시스템(recommendation system)을 구축할 수 있다. 특정 기법으로 사용자 모델링을 하였다면, 이를 이용하여 해당 사용자에게 적당한 정보를 제공하게 된다. 그 대상이 영화나 음악과 같은 멀티미디어 정보가 될 수도 있고, 쇼핑몰의 경우에는 절적한 상품정보나 카탈로그 페이지가 될 수도 있다. 이렇게 함으로써 사이트는 사용자에게 높은 만족도를 얻어내어 그 역할을 충실히 수행할 수 있고, 기업 활동의 경우에 있어서는 직접적인 매출의 신장을 기대할 수 있다. 협업 CRM의 사용자 모델링에 있어서 가장 중요한 요소는 사용자들로부터 얻어지는 피드백 정보이다. 주어진 웹 컨텐츠에 대한 사용자의 반응을 얻어내어, 사용자의 성향을 파악하고, 사용자의 특성에 맞는 상품, 정보, 페이지를 제공하는 것이다. 일반적으로 피드백은 명시적 피드백(explicit feedback)과 암시적 피드백(implicit feedback)으로 구분되며, 명시적 피드백은 컨텐츠, 상품 등에 대해 사용자로부터 직접 얻어지는 정보를 의미하고, 암시적 피드백은 마우스의 움직임, 페이지에 머문 시간, 페이지간의 이동을 나타내는 클릭 스트림(click stream)같이 사용자의 행동으로부터 간접적으로 관찰될 수 있는 정보를 말한다. 현재 구현되고 있는 대부분의 추천시스템은 명시적 피드백 중 사용자의 등급평가(rating) 정보만을 이용하고 있으며, 이것은 사용자로부터 반응을 얻기가 힘들기 때문에 데이터의 희소성 문제를 유발하고, 등급평가 정보와 같은 이산적인 데이터가 아닌 대부분의 피드백의 연속성 데이터를 다룰 수 있는 방법이 부재했다. 본 논문은 협업 CRM의 추천시스템에 적용될 수 있는 사용자모델링의 구현을 위해 웹 사이트에서 얻어질 수 있는 로그 데이터를 기반으로 사용자로부터 연속성 피드백 정보를 얻어서 통계적 학습 이론(Statistical Learning Theory: SLT)과 베이지안 퍼지 군집화(Bayesain Fuzzy Clustering: BFC) 기법을 적용 및 제안하여 추천시스템을 구축

하였다. 실제 운영되는 웹 사이트의 피드백 정보의 객관적인 데이터를 통하여 그 성능을 검증하였다.

2. 통계적 학습 이론과 베이지안 퍼지 군집화

2.1 통계적 학습 이론

본 논문의 제안 추천 시스템에 사용되는 SLT는 Vapnik(1995, 1998)에 의해 Support Vector Machine(SVM)과 Support Vector Regression(SVR)으로 체계적인 정리가 이루어 졌다. SVM과 SVR은 각각 분류(classification)와 예측(prediction)에 적용되는 모형이다.

2.1.1. Support Vector Machine

클래스 레이블들을 가진 목표변수(target variable) y와 입력벡터(input vector) x로 구성된 데이터 집합 S는 다음의 구조로 표현된다.

$$(y_1, x_1), (y_2, x_2), \cdots, (y_l, x_l), \quad x_i \in R^N, y_i \in \{-1, 1\}. \tag{1}$$

대부분의 분류모형 구축의 경우에 입력 공간(input space)에서 서로 다른 클래스 레이블을 분류하는 정확한 초평면(hyperplane)을 찾는 것은 매우 제한적이기 때문에 바로 분류 모형을 사용하기가 어렵다(Vapnik, 1995). 이 문제의 해결 방안의 하나로써 입력 공간을 더 높은 차원의 특징 공간(feature space)으로 사상(mapping)시키고, 특징 공간에서 최적의 초평면을 찾는 방법이 있다. $z=\psi(x)$를 입력 공간 R^N에서 특징 공간 Z로의 사상 ψ를 갖는 특징 공간벡터로 표현하려면 (w,b)의 쌍으로 이루어진 다음의 초평면 식을 구해야 한다.

$$w \cdot z + b = 0. \tag{2}$$

(2)식이 결정되면 (3)의 함수식에 의해 개개의 $x_i (i=1,\cdots l)$를 분류할 수 있다.

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases}. \tag{3}$$

여기서, $w \in Z$이고 $b \in R$이다. 특히, S는 (w,b)의 쌍이 존재하면 선형분류 가능(linearly separable)이라 하고, 다음의 부등식이 S의 모든 원소들에 대해 성립한다.

$$\begin{cases} (w \cdot z_i + b) \geq 1, & \text{if } y_i = 1 \\ (w \cdot z_i + b) \leq -1, & \text{if } y_i = -1 \end{cases} i = 1, 2, \cdots, l. \tag{4}$$

선형분류 가능한 집합 S에서는 두 개의 서로 다른 클래스 레이블들의 학습 데이터의 사영

(projection)들 사이의 마진(margin)을 최대화하는 유일한 최적 초평면을 구할 수 있다. 만약 S가 선형분류 가능이 아니면 분류규칙 위반(classification violations)이 SVM에서 허용되어져야 한다 (Pontil at. al., 1997). 선형분류 가능이 아닌 데이터를 다루기 위하여 음이 아닌 변수 ξ_i를 도입하여 아래 식과 같이 (4)식을 일반화한다.

$$y_i(w \cdot Z_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l. \tag{5}$$

(5)식에서 ξ_i는 (4)식을 만족하는 x_i들이다. 그러므로 $\sum_{i=1}^{l} \xi_i$는 오분류(misclassification)의 양을 나타내는 측도로 고려된다. 따라서 최적 초평면을 구하는 것은 다음 식의 문제에 대한 해가 된다.

$$\begin{array}{l} minimize \ \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \\ subject\ to\ y_i(w \cdot z_i + b) \geq 1 - \xi_i \end{array} . \tag{6}$$

여기서, $\xi_i \geq 0$이고 $i = 1, 2, ..., l$이다. C는 조정 모수(regularization parameter)인 상수(constant)이다. 이 모수의 조정으로 마진 최대화와 분류규칙 위반 사이의 균형을 맞출 수 있게 된다(Cortes et al., 1995, Pontil et al., 1997, Vapnik, 1998). (6)식에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 구할 수 있다.

$$\begin{array}{l} maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ subject\ to \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \leq \alpha_i \leq C,\ i = 1, 2, \cdots, l \end{array} . \tag{7}$$

여기서 $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$는 (5)식의 제한 조건과 관련된 음이 아닌 라그랑지 승수(multiplier)들의 벡터이다. Kuhn-Tucker 정리(Kuhn, et al., 1951)는 SVM 이론에서 중요한 역할을 한다. 이 정리에 의하여 (7)식의 해 $\bar{\alpha}_i$는 다음 식들을 만족한다.

$$\begin{array}{l} \bar{\alpha}_i (y_i (\bar{w} \cdot z_i + \bar{b}) - 1 + \bar{\xi}_i) = 0 \\ (C - \bar{\alpha}_i) \bar{\xi}_i = 0,\ i = 1, 2, \cdots, l \end{array} . \tag{8}$$

(8)식의 첫 번째 하위 식으로부터 구한 해 $\bar{\alpha}_i$는 (5)식의 등호를 만족시킨다. $\bar{\alpha}_i > 0$인 x_i를 support vector라고 부른다. 분류가 가능하지 않은 경우에는 support vector는 두 가지의 형태로 존재한다. $0 < \bar{\alpha}_i < C$인 경우, support vector x_i는 $y_i(\bar{w} \cdot z_i + \bar{b}) = 1$과 $\xi_i = 0$을 만족하고, $\bar{\alpha}_i = C$인 경우, ξ_i는 널(null)이 아니고 (4)식을 만족하지 않는 대응되는 support vector x_i는 오차(error)가 된다. $\bar{\alpha}_i = 0$에 대응되는 x_i는 결정 마진(decision margin)과 떨어져서 정확하게 분류된다. 최적 초

평면인 $\bar{w} \cdot z + \bar{b}$를 구축하기 위하여 다음 식과 $\bar{b}$가 필요하다.

$$\bar{w} = \sum_{i=1}^{l} \bar{\alpha}_i y_i z_i \ . \tag{9}$$

(9)식은 (8)식의 첫 번째 하위 식의 Kuhn-Tucker 조건에 의해 결정된다. 결정 함수(decision function)는 (3)식과 (9)식에 의해 다음 식과 같이 일반화된다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b\right). \tag{10}$$

ψ에 대한 어떠한 정보도 없기 때문에 (7)식과 (10)식의 계산은 불가능하지만 SVM에서는 ψ에 대해서 알고 있을 필요는 없다. 단지 커널(kernel)이라 불리우는 $K(\cdot, \cdot)$가 다음 식에 의해 특징 공간 Z의 내적(dot product)을 계산한다.

$$z_i \cdot z_j = \psi(x_i) \cdot \psi(x_j) = K(x_i, x_j). \tag{11}$$

Mercer의 정리(vapnik, 1998)를 만족하는 함수들은 내적 계산이 가능한 커널함수이다. SVM 분류기(classifier)를 구축하기 위하여 아래와 같은 차수(degree) d의 다항(polynimial) 커널을 사용한다.

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d. \tag{12}$$

따라서 비선형 분류 가능 초평면은 다음 식의 해로써 구해진다.

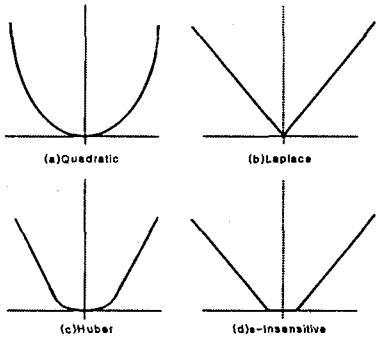
$$\begin{aligned} &maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &subject\ to \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \le \alpha_i \le C,\ i = 1, 2, \cdots, l \end{aligned} \tag{13}$$

최종적인 결정 함수는 다음 식과 같다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right). \tag{14}$$

2.1.2. Support Vector Regression

SLT중에는 특히 예측 모형에 적용되어 좋은 성능을 제공해 주는 SVR 기법이 있다. SVR은 학습 모형의 손실함수(loss function)에 대한 변화를 통하여 회귀 모형 문제에 적용될 수 있다(Smola, 1996). (그림 1)은 4가지의 손실 함수의 형태를 보여주고 있다.



(그림 1) 4가지 형태의 손실함수

(그림 1)에서 (a)는 전통적인 최소제곱 오차 기준(least square error criterion)이다. (b)는 (a)보다 이상치(outlier)에 덜 민감한 Laplacian 손실함수이다. Huber가 제안한 손실함수인 (c)는 주어진 데이터의 분포를 모를 때 최적 특성들(optimal properties)을 보여주는 robust 손실함수이다. 이러한 3개의 손실함수는 support vector들 내에 희소성(sparseness)이 있을 때에는 적당하지 않다. 이 문제를 해결하기 위하여 Vapnik(1995)은 (d)의 손실함수를 제안하였다. (d)는 Huber의 손실함수의 근사화로서 얻어진 support vector들이 희소한 집합(sparse set)으로 구성되어 있어도 사용이 가능하다. SVR도 SVM과 같은 방법을 통해 최종 해를 구한다. 특히 본 논문에서는 협업 CRM 추천 시스템의 구축에 있어서 웹 데이터의 희소성(sparseness) 문제를 해결하기 위하여 ε-insensitive 손실함수를 사용하였다(전성해 등, 2003).

2.2. 베이지안 퍼지 군집화

SVR에 의해 희소성이 제거된 웹 로그 데이터로부터 협업 CRM의 추천 시스템 구축을 위한 규칙을 생성하기 위하여 본 논문에서는 BFC를 제안하였다.

2.2.1. 군집화를 위한 퍼지 시스템 구조

퍼지 군집화에서는 군집화를 위한 유사도 정보를 가지는 분할 행렬 U를 구한다. U의 각 원소인 u_{ik}는 개체 i가 집단 k에 속하게 될 멤버 함수값을 나타낸다(Bezdek, 1987). 일반적으로 u_{ik}는 다음의 조건식을 만족한다.

$$u_{ik} \in [0,1],\ \sum_{i=1}^{K} u_{ik} = 1. \tag{15}$$

즉, 한 개의 개체에 대하여 모든 가능한 군집에 대한 소속 가능도의 합은 1이 된다. 퍼지 C-평균(Fuzzy C-Means: FCM) 방법도 퍼지 군집화 기법 중의 하나이다(Zimmermann, 2001). FCM은 (16)식의 가중 급내의 등급 제곱합(weighted within-class sum of square)을 최소화하여 군집화를 수행한다(Hathaway, et al. 1993).

$$J(U, v_1, \cdots, v_K) = \sum_{i=1}^{n}\sum_{k=1}^{K}(u_{ik})^m d^2(x_i, v_k). \quad (16)$$

(16)식에서 $v_k = (v_{ka})((k = 1, \cdots, K), (a = 1, \cdots, p))$는 집단 k의 중심값을 나타내고, $x_i = (x_{ia})$ $(i = 1, \cdots, n)$는 i번째 개체를 나타낸다. a는 입력벡터의 차원이다. $d^2(x_i, v_k)$는 x_i와 v_k간의 유클리디안 거리(Euclidean distance)를 나타낸다. m은 1에서 ∞까지의 값을 가지며 군집화의 퍼지화(fuzziness) 정도를 결정한다(Zimmermann, 2001). 즉, (16)식을 최소화하는 U와 $v_1, \ldots, v_k$를 결정하여 주어진 학습 데이터를 군집화한다.

2.2.2. 베이지안 학습을 통한 군집화 퍼지 규칙의 추출

학습 데이터(training data)의 각 개체가 특정 군집에 속할 퍼지 멤버함수를 나타내는 퍼지 군집화의 분할 행렬 U의 각 원소는 (15)식으로부터 확률과 같은 구조를 갖게 됨을 알 수 있다. 본 논문에서는 주어진 데이터로부터 베이지안 학습을 통하여 최종 사후(posterior) 확률분포로서 퍼지 군집화를 위한 분할행렬 U를 결정하였다. 퍼지 군집화를 위한 베이지안 학습에 사용되는 데이터 구조는 다음 식과 같다. (17)식은 전체 n개의 데이터 중에서 i번째 데이터에 대한 구조를 나타내고 있다(Liu, et al. 2003).

$$x_1^{(i)}, \cdots, x_{N_i}^{(i)} \sim iid\ sample\ from\ \pi_i = N(\theta_i, \Sigma_i). \quad (17)$$

즉, $x_1^{(i)}, \cdots, x_{N_i}^{(i)}$는 π_i 분포를 따르는 집단 i로부터 추출된 N_i개의 표본 데이터라고 가정한다. 위 식에서 'i.i.d.(independent, identical distributed) sample'은 임의표본(random sample)을 의미한다. π_i는 평균벡터(mean vector) θ_i와 분산-공분산행렬(variance- covariance matrix) Σ_i를 갖는 가우시안분포(Gaussian distribution)라고 가정한다. (17)식의 데이터 구조로부터 각 집단의 사전(prior) 확률분포도 역시 가우시안분포로 결정하였다(Gelman, et al., 1995). 이는 베이지안 학습의 사후 확률분포의 계산을 쉽게 할 수 있는 공액분포(conjugate distribution)의 특성을 이용하기 위함이다. 만약 공액확률분포를 사용하지 않는다면 확률적 모의실험을 통하여 사후 확률분포를 계산해야 하는 마코프 체인 몬테 칼로(Markov Chain Monte Carlo: MCMC) 기법을 사용해야 한다(Robert, et al., 1999). MCM 알고리즘은 매우 많은 계산 비용(computing cost)을 요구한다(Bishop, 1998; Press, 1989).

주어진 학습 데이터에 대한 분포인 우도함수(likelihood function)도 마찬가지 이유로 가우시안분포로 결정하였다. 따라서 군집화의 최종 U의 원소인 퍼지 멤버함수를 결정하기 위한 사후 확률분포의 구조도 가우시안분포가 된다. 다음은 베이지안 학습을 통한 퍼지 군집화의 분할행렬 U의 원소인 퍼지 규칙을 생성하는 알고리즘이다.

2.2.3. Bayesian Learning based Fuzzy Rule Extraction algorithm: Clustering approach

(단계 1) 분포의 결정

사전 확률분포의 결정: $N(\theta_i, \Sigma_i)$

~ Conjugate distribution(Gaussian)

학습 데이터의 우도함수 결정: $(x|\pi_i)$

~ Gaussian distribution

(단계 2) 사후 확률분포의 계산

Bayes' Theorem 이용

Posterior $\propto$ Likelihood$*$Prior

$$p(x \in \pi_i|x) = \frac{p(x \in \pi_i)p(x|\bar{x}_i, \Sigma_i^{-1}, \pi_i)}{\sum_{j=1}^{K} p(x|\bar{x}_j, \Sigma_i^{-1}, \pi_j)p(x \in \pi_j)}$$

~ Gaussian distribution

(단계 3) 군집화를 위한 최종 퍼지 규칙의 결정

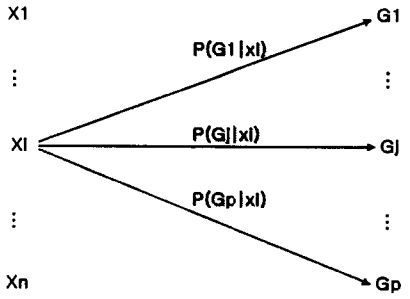
$u_{ik} = p(x \in \pi_i|x)$.

최종적으로 데이터에 대한 군집화는 다음 식과 같이 각 개체에 대한 최대 멤버함수 값을 갖는 군집으로 결정한다.

$$max_{i \in \{1,2,\cdots,K\}} p(x \in \pi_i|x). \tag{18}$$

(18)식은 확률 구조이기 때문에 퍼지 군집화의 조건인 (15)식을 만족하게 된다. 따라서 베이지안 학습을 통하여 퍼지 군집화를 위한 유사도 분할 행렬인 U를 구하였다.

다음 그림은 제안하는 알고리즘에 의해 개체가 군집에 할당되는 과정을 도식화하였다.



(그림 2) 사후 확률에 의한 학습 데이터의 군집화

(그림 2)에 의하면 개체 x_i는 다음의 식을 만족하는 집단 G^*에 할당한다.

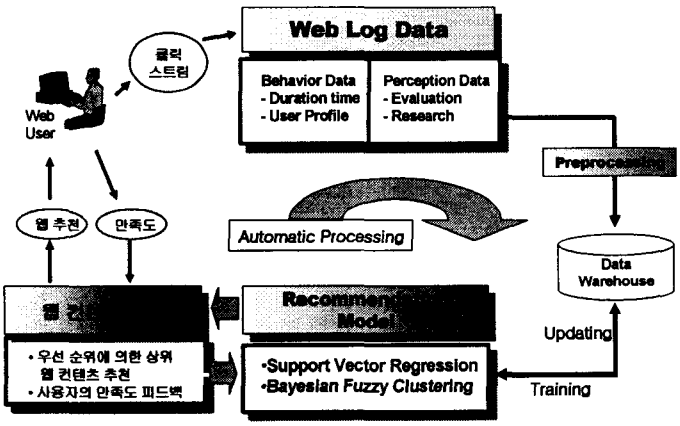
$$P(G^*|x_i) = max_{j \in \{1,2,\cdots,p\}} P(G_j|x_i). \quad (19)$$

각 개체에 대한 각 군집의 사후 확률값을 계산하여 가장 큰 값을 갖는 군집에 해당 개체를 할당하는 (19)식의 군집 판정 기준을 이용하여 퍼지 군집화를 수행한다(한진우 등, 2003).

3. 제안 추천 시스템 구축 실례

3.1. 제안 추천 시스템 구조

(그림 3)은 제안 시스템의 전체 구조이다. 우선 웹 사이트에 접속한 사용자가 현재 머물고 있는 웹 페이지들에 대한 정보가 웹 서버의 로그 파일에 저장된다. 웹 로그 파일의 클릭 스트림 데이터에는 사용자가 관심 있게 찾아보는 웹 페이지들의 일련의 순서 또는 각 페이지에서 머문 시간(duration time) 등이 저장되어 있다. 예측과 군집화를 수행하는 학습 그룹(training group)과 웹 정보의 추천을 위한 지식베이스(knowledge base), 그리고 웹 정보의 추천을 받고 이를 사용자에게 보여주는 추천 모듈(recommendation module)로 구성된다.



(그림 3) 협업 CRM을 위한 추천 시스템 구성도

(그림 3)의 협업 CRM을 위한 추천 시스템 구성도 내의 추천 모형은 SVR과 BFC의 2개의 학습 알고리즘으로 구성되어 있다. SVR 부분은 정제된 웹 로그의 클릭 스트림 데이터를 이용하여 사용자의 선호도를 예측하는 SLT 기반의 예측 모형이다. SVR을 이용하여 웹 로그의 희소성 문제를 해결하였을 뿐만 아니라, 빠른 학습 시간을 모형의 특성으로 가지는 SVR의 장점으로 인하여

실시간 예측을 필요로 하는 협업 CRM의 추천 시스템의 효율적 구축이 가능하게 되었다. SVR의 예측 모듈과 함께 서로 유사한 행동 패턴을 보이는 접속 사용자들에 대한 군집화는 제안하는 추천 시스템에서 웹 정보의 정확한 추천을 위하여 우선적으로 필요한 작업이다. 이를 위해 본 논문에서는 BFC 기법을 제안하여 SVR의 결과에 따른 웹 사용자에 대한 군집화를 수행하였다. 따라서 제안하는 시스템의 추천 모형은 SVR과 BFC가 상호 결합하여 각 사용자에 대한 웹 정보 추천을 수행하게 된다. 데이터 웨어하우스의 지식베이스는 기존의 데이터와 이를 통한 SVR과 BFC의 학습 결과에 대한 정보를 갖고 있으며, 빠른 정보 예측을 위하여 현재의 결과를 추천 모듈에 제공한다. 새로운 사용자에 대한 학습 결과에 의해서도 데이터 웨어하우스의 내용은 계속적으로 갱신된다. 사용자들에 대한 그룹화의 결과는 제안 시스템의 적용 대상이 되고, 최종적으로 사용자 그룹의 피드백 정보를 통하여 추천 시스템의 적응적 성능 향상을 가져오게 된다. 본 논문의 제안 시스템의 성능 평가에 사용할 데이터는 실제 인터넷 쇼핑몰인 Gazelle.com의 웹 로그 파일을 사용하였다. 해당 쇼핑몰 사이트는 Leg Care, Leg Ware등의 의료 장비를 취급하는 회사이다. 이 데이터는 2000년도 KDD(Knowledge Discovery & Data mining) Cup에서 경진대회 문제로 사용되었다 (KDD Cup, 2000).

3.2. 추천 컨텐츠의 예측

웹 로그 파일의 클릭 스트림 데이터는 매우 희소한 구조를 갖게 되는데, 그 이유는 웹 사이트를 구성하고 있는 전체 웹 문서 중에서 한 번의 방문을 통하여 사용자가 보게 되는 웹 페이지의 수가 상대적으로 매우 적기 때문에 발생된다. (그림 4)의 (a)와 같은 데이터 구조를 띠게 된다. 그림에서 행은 각 사용자를 나타내고, 열은 각 상품 정보를 나타내는 웹 페이지이다. 즉 User1의 사용자는 Page1은 방문한 적이 없으며 Page2는 방문하여 8초 동안 머물렀던 것이다. 하지만 전체 N개의 웹 페이지 중에서 User1은 방문한 웹 페이지의 수보다 상대적으로 방문하지 않은 웹 페이지가 훨씬 많게 된다. 때문에 웹 로그 데이터의 희소성 문제가 발생된다. 본 논문에서는 SVR 기법을 이용하여 이 문제를 해결하였다.
즉, 다음 식과 같이 i번째 웹 페이지에서 머문 시간을 i번째 웹 페이지를 제외한 전체 웹 페이지들로 모형화한다.

$$t_{page(i)} = F_{SVR}(t_{page1}, \cdots, t_{page(i-1)}, t_{page(i+1)}, \cdots, t_{pageN}). \tag{20}$$

(20)식에서 $t_{page(i)}$는 목표변수(target variable)가 되고 $(N-1)$개의 나머지 변수들이 입력벡터(input vector), $\{t_{page1}, \cdots, t_{page(i-1)}, t_{page(i+1)}, \cdots, t_{pageN}\}$가 된다. ε-insensitive 손실함수를 사용한 SVR 모형은 입력변수들 중에서 결측값이 있어도 목표변수 값을 예측할 수 있기 때문에 희소한 웹 로그 데이터의 분석이 가능하다. 이와 같은 방법에 의해 최종적으로 모든 사용자에 대한 전체 웹 페이지의 머문 시간을 예측하여 (그림 4)의 (a)의 결측 셀(missing cell)들을 채워 넣게 되면 (그림 4)의 (b)와 같은 완전한 데이터 구조가 된다.

	Page1	Page2	Page3	...	PageN
User1		8	17	...	
User2	6			...	
User3		5		...	
User4	11			...	3
User5			21	...	
⋮	⋮	⋮	⋮	⋮	
UserM		7		...	

(a)

	Page1	Page2	Page3	...	PageN
User1	*8*	8	17	...	*3*
User2	6	*9*	*13*	...	*2*
User3	*10*	5	*11*	...	*1*
User4	11	*6*	*10*	...	3
User5	*9*	*4*	21	...	*4*
⋮	⋮	⋮	⋮	⋮	
UserM	*6*	7	*12*	...	*3*

(b)

(9그림 4) 정제된 웹 로그 데이터의 구조

제안된 시스템의 성능 평가에 있어서, 각 군집의 분산 측도인 VC(variance criterion) 측도(measure)를 이용하여, 우선 BFC의 군집 결과에 대해서 기존의 계층적 군집화, K-means, 자기조직화지도(Self Organizing Maps: SOM)와 성능 평가를 수행하였다(박민재, 등, 2003). 다음으로 제안하는 추천 모형인 SVR-BFC를 기존의 피어슨의 상관 모형(Pearson's correlation)과 일반적인 연관성에 기반한 협업 추천 모형(Collaborative Filtering: CF)에 대해서 예측의 정확성을 비교하였다. 피어슨의 추천 모형은 Paul(1994)의 GroupLens 프로젝트에 기반한 방법이다. 이 모형에서 현재 사용자 a와 기존의 사용자 i와의 연관도를 나타내는 가중치인 $w(a,i)$를 다음 식으로 정의한다.

$$w(a,i)=\frac{\sum_j (v_{aj}-\overline{v_a})(v_{ij}-\overline{v_i})}{\sqrt{\sum_j (v_{aj}-\overline{v_a})^2}\sqrt{\sum_j (v_{ij}-\overline{v_i})^2}}. \tag{21}$$

여기서 v_{ij}는 사용자 i가 아이템 j를 선택(rating)한 것을 나타내고 $\overline{v_a}$는 사용자 a가 평균적으로 아이템을 선택한 것이다. 피어슨의 추천모형과 함께 본 논문의 추천 모형과 비교되는 또 다른 협

업 추천 모형은 코사인 유사도(cosine similarity)를 이용하는 Amazon(Linden, 2003)의 협업 추천 기법이다.

3.3 군집화의 성능 평가

(그림 4)의 (b)처럼 SVR을 이용하여 KDD Cup 2000의 정제된 로그 파일로부터 데이터의 희소성을 제거한 후 제안 모형과 비교 모형들의 군집 결과에 대한 성능 평가 실험을 수행하였다. 군집화의 성능 평가를 위한 측도로는 앞에서 언급한 VC를 이용하였다. VC에서 결정된 군집 결과에 대한 성능평가 기준에는 각 군집의 평균 밀도(average density)와 군집수 증가에 대한 불이익(penalty)의 두 조건을 포함하고 있다. 이러한 VC 측도는 다음 식과 같이 정의된다(박민재 등, 2003).

$$VC_M = \frac{\sum_{i=1}^{M} V_i}{M} + 0.1M \tag{22}$$

(22)식에서 M은 군집의 수를, V_i는 i번째 군집의 분산을 의미한다. (22)식의 첫 번째 항은 각 군집의 평균밀도를 의미하고, 두 번째 항은 군집수 증가에 따른 불이익을 나타낸다. 따라서 (22)식은 서로 동질적인 것들끼리 더 잘 묶여지는 좋은 성능의 군집에 대해서 보다 작은 값을 갖게 된다. 두 번째 항의 균형 상수는 명확한 군집의 구분을 가지는 인공 데이터와 잘 알려진 기계 학습 데이터를 이용한 실험을 통해 0.1로 결정되었다(UCI). 제안 모형을 포함하여 본 논문에서 비교되는 4가지 군집화 모형의 VC 결과값은 다음 표와 같다.

<표 2> 4가지 비교 모형들의 VC 값

군집수	계층적군집화	K-Means	SOM	BFC
3	0.87	0.98	0.72	0.65
4	0.84	0.62	0.68	0.55
5	0.79	0.63	0.64	0.58
6	0.88	0.72	0.69	0.59
7	0.89	0.94	0.68	0.65
평균	0.85	0.78	0.68	0.60

다른 3개의 모형에 비해 제안하는 BFC 기법에 의한 VC값의 평균값이 가장 작게 나왔다. 즉, 다른 군집화 기법에 비해 BFC가 더 유사한 사용패턴을 보이는 사용자들끼리 잘 그룹화한 것이다. 특히 BFC에 의한 군집화 결과 중에서도 4개의 군집수로 군집화를 수행했을 때의 VC값이 가장 작게 나왔다. 이 때의 군집결과에 대한 각 군집의 동질성이 가장 우수하다는 것으로 해석된다. 즉, 군집화의 기본적인 전략이 군집 간의 분산은 최대로 하고 군집 내의 분산은 최소로 하는 전략과 같은 개념이다.

3.4 추천 시스템을 위한 BFC 결과

<표 2>의 결과에 의해 군집화를 위한 4가지 비교 모형들 간의 군집 결과, 최적 군집 수는 BFC의 4임을 알 수 있었다. 따라서, 4개의 군집수로 최종 BFC 군집 분석을 수행한 결과로부터 다음 표를 구할 수 있었다.

<표 3> BFC를 이용한 군집 결과

군집	선호도가 높은 상위 3개의 웹 문서(through duration time)		
	1st	2nd	3rd
G1	page 29 (11.3)	page 43 (9.6)	page 36 (6.8)
G2	page 126 (8.4)	page 138 (6.3)	page 92 (4.2)
G3	page 56 (13.2)	page 55 (11.9)	page 46 (8.5)
G4	page 11 (16.6)	page 39 (10.3)	page 52 (6.2)

위의 결과는 각 군집에 최종적으로 할당된 사용자들의 정보를 이용하여 해당 군집내의 사용자들이 머물 시간이 가장 높을 것으로 예측되는 상위 3개의 웹 문서를 보여 준다. 각 셀은 각 군집의 해당 순위에 속하는 웹 문서와 머문 시간을 나타내고 있다. 예를 들어, 그룹 1(G1)에 속한 사용자들은 가장 우선순위가 높은 웹 페이지로 page29가 되고 이 페이지에 머물게 될 예측 시간은 11.3초가 된다. 즉, 그룹 1에 속하는 사용자들은 이 페이지에 대한 정보가 우선적으로 필요한 것이다. 만약 새로운 사용자가 군집 1에 속하게 된다면 이 사용자에게는 웹 페이지 29를 우선적으로 추천하게 된다. 다음으로는 page43과 page36의 순으로 추천하게 된다. 총 13,109개의 세션(session) 사용자들 중에서 8,652 개의 세션들이 협업 CRM을 위한 추천 시스템 구축을 위한 학습 데이터로 사용되었다. 나머지 4,457개의 세션들을 테스트 데이터로 사용하여 제안 시스템의 성능 평가를 수행하였다. 테스트 데이터의 각 사용자는 4개의 군집들 중에서 입력 변수들에 의한 유클리디안 거리를 이용하여 가장 유사한 군집이 결정되었다. 본 논문의 제안 시스템의 예측의 정확도는 최소제곱평균(mean square error: MSE)을 이용하였다. MSE는 추천한 값과 실제 값의 차이의 제곱 평균이다. MSE 측도를 이용하여 제안 모형과 현재 가장 많이 사용되는 피어슨 기반의 추천 모형과 일반적인 협업 추천 모형을 비교하였다.

<표 4> 최종 성능 평가 결과

	피어슨 시스템	CF	SVR-BFC
MSE	1.66	1.32	0.79

<표 4>를 통해 SVR를 통한 BFC에 의한 제안 추천 기법의 MSE값이 기존 피어슨 기반의 추천 모형이나 일반적인 협업 추천 모형에 비해 작음을 알 수 있었다. 이는 본 논문에서 제안하는 협업 CRM을 위한 웹 컨텐츠 추천 시스템이 기존의 것에 비해 좀 더 정확한 예측 결과를 보이고

있음을 알 수 있었다.

4. 결론

본 논문에서는 SVR 기법에 의해 분석 데이터로 사용되어지는 웹 로그 파일의 클릭 스트림 데이터가 가지는 희소성 문제를 해결하여 완전한 데이터 구조를 이루게 된 웹 로그에 대하여 제안한 BFC를 이용하여 추천 시스템에 적용할 모형을 만들었다. 즉, 사용자가 웹 사이트에 접속하여 각 페이지에 머문 시간을 SVR 기법을 이용하여 각 페이지 별로 모형화 하였고, 총 269개의 웹 페이지를 가진 실험 데이터에 대한 예측 모형을 통해 사용자 별로 아직 접속하지 않은 웹 페이지에 대한 접속 가능 시간을 예측하여 가장 접속 시간이 클 것으로 예상되는 페이지를 해당 사용자의 선호 웹 정보로 추천하는 웹 페이지 사용자 모델링 방법을 제안하였다. 제안 방법이 기존의 것에 비해 예측의 정확성이 향상되었음이 실험을 통하여 확인되었다. 향후에도, Import Vector Machine과 같은 다양한 통계적 학습 이론을 적용하여 실시간으로 변화되는 인터넷 환경에 적응적으로 대처할 수 있는 더욱 지능화된 추천 시스템을 개발할 수 있을 것이다.

참고문헌

[1] 박민재, 전성해, 오경환, (2003) 붓스트랩 기법과 유전자 알고리즘을 이용한 최적 군집수 결정, 퍼지 및 지능시스템학회 논문지 제13권, 제1호, pp. 12-17.

[2] 전성해, 임민택, 전홍석, 황진수, 최성용, 김지연, 오경환, (2003) Web Log Analysis using Support Vector Regression, 한국통계학회논문집, 제10권, 제1호, pp. 61-67.

[3] 한진우, 전성해, 오경환, (2003) 군집화를 위한 베이지안 학습 기반의 퍼지규칙 추출, 한국정보과학회 2003 춘계학술대회 발표논문집(II).

[4] Basu C., Hirsh H., Cohen W., (1988) Recommendation as classification : Using Social and Content-based Information in Recommendation, Proceedings of the Workshop on Recommendation system. AAAI Press, Menlo Park California.

[5] Bezdek J. C., (1987) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press.

[6] Bishop C. M., (1998) Neural Networks for Pattern Recognition, Clarendon Press:Oxford.

[7] Cortes C., Vapnik V. N., (1995) Support Vector Networks, Machine Learning, vol. 20, pp. 273-297.

[8] Fisher D., Hildrum K., Hong J., Newman M., Thomas M., Vuduc R., (2000) SWAMI: A Framework for Collaborative Filtering Algorithm Development and Evaluation, SIGIR 2000, Athens, Greece.

[9] Forrester Research, (2002) http://www.forrester.com.

[10] Gelman A., Carlin J. B., Stern H. S., Rubin D. B., (1995) Bayesian Data Analysis, Chapman & Hall.

[11] Hathaway R. J., Bezdek J. C., (1993) Switching Regression Models and Fuzzy Clustering,

IEEE Trans. Fuzzy System, Vol. 1, No. 3, pp. 195-204.
[12] KDD Cup, www.ecn.purdue.edu/KDDCUP.
[13] Kuhn H. W., Tucker A. W., (1951) Nonlinear programming, In proceeding 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics, pp. 481-492.
[14] Linden, G., Smith, B., York, J., (2003) Amazon.com recommendations: item-to-item collaborative filtering, IEEE Internet Computing, Vol. 7, Issue 1, pp. 76 - 80
[15] Liu J. S., Zhang J. L., Palumbo M. L., Lawrence C. E., (2003) Bayesian Clustering with Variable and Transformation Selections, Bayesian Statistics 7, Oxford University Press.
[16] Paul R., Neophytos L.,Mitesh S., Peter B., John R., (1994) GroupLens: An Open Architecture for Collaborative Filtering of Netnews, ACM conference on Computer Supported Cooperative Work.
[17] Pontil M., Verri A., (1997) Properties of support vector machine, MIT AI Memo, No. 1612.
[18] Press S. J., (1989) Bayesian Statistics: Principles, Models, and Applications, John Wiley & Sons.
[19] Robert C. P., Casella G., (1999) Monte Carlo Statistical Methods, Springer.
[20] Smola A. J., (1996) Regression estimation with support vector learning machines, Master's thesis, Technische University Munchen.
[21] UCI Machine Learning Repository, www.ics.uci.edu/~mlearn.
[22] Vapnik V. N., (1995) The Nature of Statistical Learning Theory, New York: Springer-Verlag.
[23] Vapnik V. N., (1998) Statistical Learning Theory, Wiley, N.Y.
[24] Zimmermann H. J., (2001) Fuzzy Set Theory and Its Application, Kluwer Academic Publishers Group.

[2003년 9월 접수, 2004년 4월 채택]