MCMC 결측치 대체와 주성분 산점도 기반의 SOM을 이용한 희소한 웹 데이터 분석

전 성 해† · 오 경 환††

요 약

웹으로부터 유용한 정보를 얻기 위한 연구는 현재 많이 진행되고 있다. 본 논문에서는 특히 웹 로그 데이터의 희소성에 대한 문제 해결과 이를 통한 웹 사용자의 군집화 방안에 대하여 연구하였다. MCMC 방법의 베이지안 추론에 의한 결측치 대체 기법을 이용하여 웹 데이터의 희소성을 제거하였고, 주성분에 의한 산점도를 통하여 형상지도의 차원을 결정한 자기 조직화지도를 이용하여 웹 사용자의 군집화를 수행하였다. 제안 기법은 기존의 방법들에 비해 모형의 정확도와 빠른 학습 시간을 제공하여 주었다. KDD Cup 데이터를 이용한 실험을 통하여 제안 방법에 대한 문제 해결 절차 및 성능 평가를 객관적으로 확인하였다.

Sparse Web Data Analysis Using MCMC Missing Value Imputation and PCA Plot-based SOM

Sung-Hae Jun† · Kyung-Whan Oh††

ABSTRACT

The knowledge discovery from web has been studied in many researches. There are some difficulties using web log for training data on efficient information predictive models. In this paper, we studied on the method to eliminate sparseness from web log data and to perform web user clustering. Using missing value imputation by Bayesian inference of MCMC, the sparseness of web data is removed. And web user clustering is performed using self organizing maps based on 3-D plot by principal component. Finally, using KDD Cup data, our experimental results were shown the problem solving process and the performance evaluation.

키워드 : 변형된 MCMC 결측치 대체(Hybrid MCMC Missing Value Imputation), 자기 조직화지도(Self Organizing Maps), 주성분 산점도(Principal Component based Plot)

1. 서 론

인터넷에서 유용한 정보를 얻기 위한 연구는 현재 많이 진행되고 있다. 이 때 분석에 주로 사용되는 웹 로그의 클릭 스트림 데이터는 전체적인 데이터의 테이블 구조에서 각 셀에 대한 결측치가 매우 많이 때문에 효과적인 정보 예측 모형을 위한 학습 데이터로서 직접 사용하는데는 어려움이 있다. 클릭 스트림 데이터 테이블의 각 셀에 결측치가 많아 웹 로그 데이터의 희소성이 발생하는 이유는 웹 사이트의 전체 페이지 중에서 각 사용자가 한 번의 접속으로 보게되는 웹 페이지의 수가 상대적으로 매우 작기 때문이다. 본 논문에서는 이러한 웹 로그 데이터의 희소성(sparseness)에 대한 문제 해결 방안에 대하여 연구하였다. 또한 웹 사용자들을 서로 유사한 행위 패턴을 보이는 사용자들 끼리 군집화 하는 방법에 대한 연구도 동시에 수행하였다. 마코프 연쇄 몬테 칼로(markov chain monte carlo : MCMC) 방법의 베이지안 추론을 적용한 결측치 대체 기법(missing value imputation)을 이용하여 웹 데이터의 희소성을 제거하였고 이 결과로서 얻게 된 완전한 클릭 스트림 데이터에 대한 주성분 분석을 통하여 보유 주성분에 대한 산점도를 시각적으로 관찰하여 형상 지도(feature maps)의 차원을 결정한 자기 조직화 지도를 이용하였다. 2장에서는 희소한 웹 데이터 분석을 위한 기존의 방법과 문제점에 대해서 살펴보고, 3장에서는 본 논문에서 제안하는 결측치 대체 기법과 다변량 통계 분석 기법 중 하나인 주성분 산점도에 기반한 자기 조직화 지도의 군집화 전략을 설명하였다. KDD-Cup 2000 데이터를 이용한 실험을 통하여 제안 방법에 대한 문제 해결 절차 및 모형에 대한 성능 평가는 4장에서 논의하였고, 마지막으로 5장에서는 결론 및 향후 연구과제에 대하여 알아보았다.

† 정 회 원 : 청주대학교 통계학과 전임강사
†† 정 회 원 : 서강대학교 컴퓨터학과 교수
논문접수 : 2002년 9월 30일, 심사완료 : 2002년 12월 20일

2. 기존의 웹 데이터 분석의 문제점

웹 로그 데이터의 분석을 통하여 사용자에 대한 추천 시스템과 같은 정보 서비스를 하기 위해서는 매우 많은 트랜젝션 데이터가 필요하다[1]. 이러한 데이터의 상당 부분은 웹 서버의 로그 파일에 저장되어 있고 이러한 데이터는 매우 희소한(sparse) 데이터 구조를 갖는다. 왜냐하면 웹 서버에 있는 수많은 웹 문서들 중에서 한 번에 특정 사용자가 접속하여 보게 되는 문서의 수는 매우 적기 때문이다. 따라서 정제된 웹 로그 데이터의 각 열(column)이 개개의 웹 페이지를 나타내고 각 행(row)은 접속한 각각의 사용자들 일 때, 이 클릭스트림 데이터는 매우 희소한 데이터 구조를 띠게 된다. 이러한 경우 특히 웹 추천 등을 위한 예측 모형 구축에서는 다양한 문제점들이 발생한다. 즉, 사이트 전체의 웹 페이지에 비해서 특정 사용자가 접속한 페이지의 수가 상대적으로 너무 적기 때문에 전체 페이지 각각에 대한 접속 가능 시간을 예측하기가 어렵게 된다. 특히 피어슨(Pearson)의 상관 계수(correlation coefficient) 알고리즘을 사용하는 선호도 예측 시스템에서는 모형에 대한 성능 저하가 나타난다[2]. 왜냐하면 이 방법은 전체 데이터에 대한 평균이 반드시 필요한데 결측치가 많게 되면 당연히 모형 구축에 사용되는 데이터 중에서 결측치를 포함한 개체는 빠지게 되어 학습 데이터의 크기가 크게 감소되면서 이 값에 대한 신뢰도가 떨어지게 되기 때문이다. 따라서 현재 이러한 문제를 해결하기 위한 방법으로는 비정치값 분해(singular value decomposition)와 같은 희소한 데이터에 대한 차원 축소를 통해 데이터 마이닝 기법을 적용하는 연구들이 진행되고 있다[12, 13]. 하지만 이러한 방법들은 차원의 축소에 의해 원래 데이터에 대한 정보의 손실이 발생하여 모형에 대한 설명력의 저하를 감수해야 한다. 본 논문에서는 차원의 축소를 하지 않고 원래 데이터의 정보를 그대로 유지하면서 결측치 대체 전략을 취하여 웹 데이터의 희소성 문제를 해결한다.

3. 제안된 희소한 웹 데이터 분석 기법

3.1 변형된 MCMC를 이용한 결측치 대체

희소한 웹 데이터를 그대로 분석하면 데이터의 부족에 따른 모형의 성능 저하가 예상된다. 모든 변수에 대해서 모든 레코드의 값이 완벽하게 채워진 경우(complete cases)에 비해 결측값을 많이 포함하고 있는 경우(incomplete cases)에는 모형에 대한 정보의 손실이 필연적으로 나타나게 된다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위한 전략으로서 다중 결측치 대체 방법을 사용하였다. 이러한 방안으로 특히 통계 물리학(statistical phisics) 분야에서 입자 데이터의 모형화에 사용되고 있는 MCMC 방법을 웹 로그 데이터의 결측치 대체에 새롭게 적용하였다. 이 방법은 현재 통계학 분야에서 전체 데이터 크기에 비해 결측치의 상대적 빈도가 적은 소규모 데이터의 결측치 처리에 사용되었다[11]. 하지만 본 논문에서 실험하는 웹 로그 데이터는 데이터의 크기가 1.2GB로서 매우 크며 또한 전체 데이터에서 결측치의 비율이 상대적으로 매우 큰 경우이다. 따라서 기존의 MCMC 결측치 대체 기법을 그대로 사용하게 되면 모형의 예측력이 크게 떨어질 뿐만 아니라 학습 시간도 매우 많이 소요된다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위하여 MCMC 대체 기법에 사전 확률 분포 선택 단계를 새롭게 추가하여 대규모의 희소 웹 데이터의 분석을 가능케 하였다.

결측치 대체 기법은 웹 로그 데이터와 같이 희소한 데이터를 분석해야 하는 경우에 모형의 성능을 유지, 향상시킬 수 있는 좋은 전략이 될 수 있다. 본 논문에서는 각 결측치에 대해서 한 개의 값을 추정하여 채워넣는 기존의 회귀모형과는 달리 결측치의 참값(right value)에 대한 불확실성을 고려하여 가능한 값들의 집합으로서 각 결측치를 대체하는 다중 결측 데이터 대체 방법(multiple missing data imputation)을 수행한다[11]. 이러한 방법의 하나로서 본 논문에서는 변형된 MCMC 방법을 제안하였다.

모든 변수에 대한 값이 완전하게 갖추어진 데이터에 대한 표준 모형과 이것을 이용한 분석 결과들의 결합에 의해 결측치에 대한 대체가 이루어진다. 결측치 대체를 위한 마코프 연쇄 몬테 칼로(Markov Chain Monte Carlo : MCMC) 기법은 원래 물리학에서 상호 작용하는 분자들의 평형 분포(equilibrium distribution)를 구하는 도구(tool)로서 사용되었다[9]. 본 논문에서는 마코프 연쇄를 통해 다차원 확률 분포(multi-dimensional probability distribution)로부터 의사난수(pseudorandom)를 생성하여 결측치를 대체하는데 사용한다. 마코프 연쇄는 과거의 모든 시점은 아무 관계가 없이 바로 앞 시간의 상태에만 의존하는 조건부 확률을 나타내는 사건들을 가리킨다. MCMC의 정상 분포(stationary distribution)를 구하여 반복된 연쇄의 모의 실험을 통하여 결측치 대체에 사용할 분포를 구한다. 이 때 사용되는 베이지안 추론에서 미지의 모수에 대한 정보는 사후(posterior) 확률 분포의 형태로 표현된다. MCMC는 베이지안 추론에서 사후 확률 분포를 구하는 방법으로써 사용된다. 즉 MCMC를 통하여 미지의 모수들에 대한 결합 사후 확률 분포를 구하고 이중 추정해야 할 모수를 사후 확률 분포를 이용한 모의 실험을 통해 구하게 된다[12]. 데이터의 분포가 다변량 정규 분포라고 가정하고 베이지안 추론에 의한 결측치 대체를 위한 데이터의 확장(augmentation)은 다음과 같은 2단계의 과정을 거친다.

[1 단계] Imputation Step (I-step)

① 추정된 평균 벡터와 공분산 행렬로부터 독립적으로 각 관측치를 위한 결측치를 구한다. (bootstrapping 사용)

② $X_{i,mis}$: i번째 관측치를 위한 결측치 변수

$X_{i,obs}$: i번째 관측치

라고 하면,

③ $X_{i,mis}$를 생성하기 위한 확률 분포, $P(X_{i,mis} \mid X_{i,obs})$를 구한다.

[2 단계] Prior Selection Step (PS-step)

① 결측치 대체 분포는 데이터에 의한 우도 함수(likelihood function)와 사전 확률 분포(prior probability distribution)의 곱에 비례한다.

$$posterior \propto likelihood \times prior$$

② 대용량 데이터의 학습 시간을 고려하여 공액 사전 확률 분포를 결정한다. ~ Gaussian prior, $f(x)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

[3 단계] Posterior Computing step (PC-step)

① I-step과 PS-step의 완전한 표본 추정치로부터 사후 모 평균 벡터와 모 공분산 행렬을 구한다.

② 모수에 대한 사전 정보가 없으면 모호 사전 확률(non-informative prior)을 사용할 수 있다.

③ 최종적으로 사후 확률 분포를 계산하여 결측치 대체 분포로 사용한다. 결측치에 대한 대체값은 이 분포로부터 생성되는 값으로 한다.

```
// Hybrid MCMC Missing value Imputation algorithm
1. With Current parameter estimate,
     θ^(t) at t^th iteration
   I-step draws X_mis^(t+1) from      P(X_mis | X_obs, θ^(t))
   PS-step and PC-step draw θ^(t+1) from
                                      P(θ^(t) | X_obs, X_mis^(t+1))
2. This create a Markov Chain
     ( X_mis^(1), θ^(1)), ( X_mis^(2), θ^(2)), ···
   which converges in distribution to  P(X_mis , θ| Y_obs)
```

(그림 1) MCMC 결측치 대체 알고리즘

I-step, PS-step, 그리고 PC-step은 신뢰할 만한 값이 나올 때까지 반복하여 수행된다[14]. 반복의 최종 목표는 데이터들이 정상 분포로 수렴되어 서로 독립적으로 결측치의 대체값을 생성할 수 있게 되어야 한다. PS-step에서 공액 사전 분포를 사용함으로써 깁스 샘플링(Gibbs sampling)이나 메트로폴리스(Metropolis)의 과정을 거치지 않게 됨으로써 MCMC의 학습 시간을 단축하는 효과를 보았다. (그림 1)은 3개의 단계를 갖는 변형된 MCMC 결측치 대체 알고리즘을 나타내고 있다.

제안 방법과 다른 결측치 대체 기법과의 성능 평가를 위하여 본 논문에서는 평균제곱오차(mean squared error : MSE)를 사용한다. 이 식은 다음과 같이 표현된다[8].

$$MSE_Y = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 / df \qquad (1)$$

식 (1)에서 n은 전체 학습 데이터의 수이다. Y_i는 실제값이고 $\widehat{Y}_i$는 예측값이다. 즉 MSE는 실제값과 예측값과의 차이에 대한 제곱합이다. 이 값이 작을수록 결측치를 정확히 추정하는 모형이 된다. df는 자유도(degree of freedom)로서 (전체 자료의 수 - 1)이 된다[10].

3.2 주성분 산점도 기반의 SOM을 이용한 사용자 군집화

주성분 분석(principal component analysis)은 다차원 변수들을 이들 간의 상관 구조를 통하여 요약하여 변수들 상호간의 복잡한 구조를 단순화하는 기법이다[3]. 즉, 변수들을 선형 변환시켜 주성분이라고 부르는 상대적으로 서로 독립적인 새로운 잠재 변수들을 유도한다. 이 때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도의 순서를 생각할 수 있는데, 주로 사용되는 판단 기준은 고유치(eigen value)와 전체 데이터에 대한 설명력을 나타내는 누적 부산이다. 본 논문에서도 이 두 가지 기준을 이용하여 주성분을 결정한다. 전체 사용자의 각 페이지에 대한 데이터를 주성분을 이용하여 변환시킨다. 이렇게 2개~3개 정도의 주성분으로 변형된 데이터는 산점도(plot)로 나타내면 시각적으로 전체 군집의 수를 파악할수 있게 된다. 더 많은 주성분을 이용하지 않는 이유는 시각적으로 인간이 판단할수 있는 차원이 3차 이내이기 때문이다. 이 정보를 이용하여 자기 조직화 지도의 형상 지도의 차원을 결정한다.

여러 가지 신경망 모형 중에서 인간의 뇌 구조를 가장 잘 모형화한 자기 조직화 지도(Self Organizing Maps : SOM)는 1980년대 초에 코호넨(Kohonen)에 의해 제안된 신경망 모형이다[5-7]. SOM은 신경망 중에서도 학습 자료에 대한 결과값을 모르고 학습이 수행되는 자율 학습(unsupervised learning)의 구조를 가지고 있다. 음성 인식, 문자 인식, 구문 분석 등 다양한 분야에 응용되는 SOM은 입력층과 출력층으로 구성된 순방향 단층 신경망 구조를 갖는다. SOM의 연결강도인 가중치는 정규화된 입력 벡터에 대응되는 출력 노드의 중심값과 같은 역할을 하며 학습 동안에 입력 벡터와 가장 가까운 측도를 갖는 출력 노드가 승자(winner)가 되고, 이 승자 노드와 이웃 하는 것들의 가중치만이 갱신한다. 특히 자기 조직화 지도는 다층 신경망(multi layer perceptron)과 같은 지도 학습(supervised learning) 모형에 비해서 매우 단순한 2개의 층(layer)으로 이루어지면서 다차원의 자료를 2차원의 형상 지도(feature maps)로 투영(projection)시켜 스스로 경쟁 학습(competitive learning)을 할 수 있도록 한다. SOM을 사용할 때는 다른 신경망 들에서는 일반적으로 필요하지 않은 작업인 연결강도 벡터와 입력 벡터를 통상 0에서 1 사이의 정규화된(normalized) 값으로 변형하는 작업을 한다. 이는 승자노드의 결정에 유클리디안 거리 측도를 사용하기 때문이다. 즉 최소 거리를 갖는 노드가 승자가 된다.

```
// SOM algorithm by T. Kohonen
[Step 1] Initialization
   Choose random values for the initial weights
   Determine the size of feature maps
[Step 2] Winner finding
   Find the winner neuron j*
```

using the minimum-distance criterion

$$j^* = \arg\min_j \| x(k) - w_j \|$$

where $x(k)$ represents input vector

[Step 3] Weight updating

Adjust the weights of the winner and its neighbors, using the following rule

$$w_j(k+1) = w_j(k) + \eta(k)(x(k) - w_j(k))$$

where $\eta(k)$ is a positive constant and $N_j^*(k)$ is the neighborhood set of the winner

(그림 2) SOM 알고리즘

(그림 2)에서 SOM 알고리즘은 초기화 단계, 승리노드 결정 단계, 그리고 가중치 갱신 단계로 이루어진다. 본 논문의 주성분 산점도는 초기화 단계에서의 형상 지도 차원 결정에 사용되었다.

4. 실험 및 결과

본 논문의 실험을 위한 웹 로그 데이터는 KDD-Cup 2000에서 문제로 주어졌던 로그 데이터로써 인터넷 쇼핑몰(Gazelle.com)의 2개월 간의 클릭스트림 만을 모아놓은 1.2GB의 텍스트 데이터이다[16]. 해당 쇼핑몰은 의료용 장비인 Leg-care 혹은 Leg-wear 제품을 전문적으로 판매하는 업체로서 데이터는 이러한 인터넷 쇼핑몰의 로그라는 특성으로 인하여 매우 방대한 양의 데이터를 가지고 있지만 (그림 3)에서 보는 것처럼 매우 희소한(sparse) 구조를 갖고 있다. 즉, 사용자 1 (User 1)은 269개의 전체 웹 페이지 중에서 8개의 페이지만 접속하였다. 또한 학습 데이터의 마지막 사용자인 13109번째 사용자(User 13109)는 단지 5개의 웹 페이지 만을 접속하였다. 그림에서 각 페이지의 괄호안의 수치는 해당 페이지에 접속한 시간(duration time)이다. 이처럼 한번의 웹 페이지를 찾은 사용자가 접속하여 보게 되는 페이지 수는 전체 웹 페이지에 비해 매우 작은 수가 된다.

User 1	Page 1 (2 seconds), Page 3 (7 seconds) Page 16 (3 seconds), Page 88 (6 seconds) Page 114 (11 seconds), Page 175 (9 seconds) Page 201 (1 second), Page 268 (3 seconds)
⋮	⋮
User 13109	Page 1 (5 seconds), Page 6 (3 seconds) Page 98 (4 seconds), Page 164 (9 seconds) Page 168 (5 seconds)

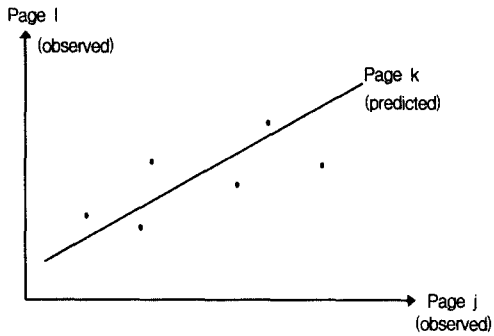
(그림 3) 희소한 웹 로그 데이터 구조

즉, 웹 로그 데이터가 희소한 구조를 갖게되는 이유는 웹 사이트 내의 수 많은 웹 페이지 중에서 한 번의 접속 사용자가 보게되는 웹 페이지의 수가 전체 문서에 비해 상대적으로 매우 작기 때문에 발생한다. 본 논문의 실험을 위한 KDD Cup 2000 데이터를 정제한 클릭스트림의 정보는 <표 1>과 같다.

<표 1> 정제된 실험 데이터 정보

구 분	데이터 범위
사 용 자	13109(명)
웹 페이지	269(개)
duration	0~1000(초)

즉, 최초의 로그 파일을 정제한 로그 데이터의 유효한 총 사용자 수는 13,109명이고 해당 사이트의 유효한 총 웹 페이지 수는 269개가 된다. 각 사용자가 해당 웹 페이지를 방문하여 머문 시간은 0에서 1000사이의 초단위 시간으로 표시하였다. 실험에서 전체 데이터의 2/3는 학습 데이터(training data)로 사용하고 나머지 1/3은 MSE를 구하기 위한 테스트 데이터(test data)로 사용한다. 웹 로그 데이터의 희소성을 제거하기 위한 모형의 구조는 (그림 4)처럼 표현된다. 각 점은 사용자 한 명에 대한 해당 페이지의 접속 시간을 나타낸다. 즉, 알고있는 i 번째 페이지와 j 번째 페이지의 접속 시간을 이용하여 k 번째 웹 페이지의 접속 가능 시간을 예측하게 된다.



(그림 4) 결측치 대체 모형 구조

따라서 그림에서 처럼 해당 페이지를 제외한 다른 페이지의 시간에 따라 특정 사용자의 해당 페이지에 관한 접속 시간을 예측하게 되며 이 값으로 결측값을 대체하게 된다. 본 논문에서는 이러한 결측치 대체 모형으로서 변형된 MCMC 방법의 베이지안 사후 확률 모형을 사용한다. 예를 들어 (그림 3)과 같이 웹 페이지 데이터 구조에서 i번째 페이지에 대한 접속 가능 시간에 대한 예측 모형은 다음과 같다.

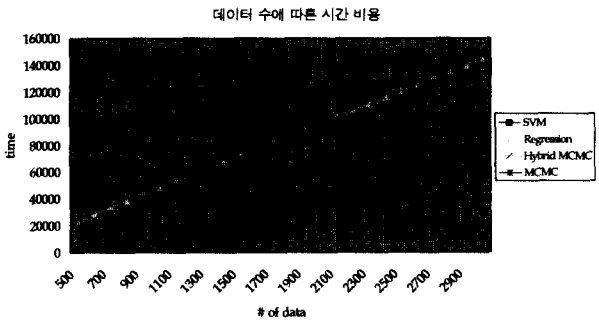
$$pref_{page_i} = f(page_1, \cdots, page_{i-1}, page_{i+1}, \cdots, page_{269}) \quad (2)$$

식 (2)를 통해 페이지 i에 대한 선호도를 i번째 페이지를 제외한 268개의 페이지를 이용하여 예측할 수 있다. 기존의 결측치 대체 모형과의 정확도에 대한 성능 평가를 한 결과가 <표 2>와 같이 나타났다. 기존의 회귀 모형(regression)에 비해서는 매우 우수한 성능을 보이고 있고 매우 빠른 학습 시간을 갖는 Support Vector Machine[14]에 비해서도 더 정확한 결측치 예측 능력을 가지고 있는 것을 알 수 있다. 또한 기존의 MCMC 대체 기법을 그대로 사용하면 예측의 정확도가 본 논문에서 제안하는 변형된 MCMC에 비해서 좋지 않음을 알수 있다.

〈표 2〉 MCMC 결측치 대체 기법의 성능 평가

방 법	MSE
Regression	0.93
SVM	0.76
MCMC	0.86
Hybrid MCMC	0.61

(그림 5)은 <표 2>에서 비교한 4개의 모형간의 학습 시간에 대한 성능 평가를 수행하였다. 그래프에서 가로축은 데이터의 수이고, 세로축은 학습시간(초)을 나타내고 있다. Support Vector Machine과 MCMC 기법이 차이를 거의 보이지 않으면서 빠른 학습 시간을 보이고 있음을 알 수 있다. 데이터의 수가 증가할수록 회귀 모형과의 성능 차이는 더욱 커짐을 알 수 있다.



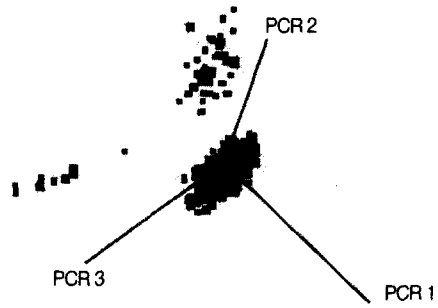
(그림 5) 모형간 학습 시간 비교 결과

또한 기존의 MCMC 방법은 제안하는 변형된 MCMC 방법에 비해 데이터의 학습 시간이 매우 많이 소요됨을 알 수 있다. SOM을 이용한 사용자 군집화를 수행할 때 우선 고려해야 하는 문제는 군집화가 수행되어지는 형상 지도의 차원을 결정하는 것이다. 이 차원의 크기를 크게 하면 학습이 끝난 후의 군집의 수가 많아지게 되고, 작게 하면 반대의 결과가 나타난다. 따라서 적절한 크기의 차원을 결정해야 하는 문제가 발생한다. 대부분의 SOM의 사용에서는 이 문제를 시행 착오(trials an errors)를 거치면서 휴리스틱하게 결정하고 있다. 본 논문에서는 이 문제를 해결하기 위하여 다변량 통계 기법 중의 하나인 주성분 분석을 이용한다. <표 3>는 정제된 KDD Cup 데이터에 대하여 주성분 분석을 수행한 결과이다.

〈표 3〉 상위 3개의 주성분의 설명력

주 성 분	고 유 값	누적비율
1	1.75	43.81(%)
2	1.20	73.98(%)
3	0.78	93.55(%)

<표 3>의 결과에 의하면 3개의 주성분만을 이용해도 데이터의 93.55%가 설명되어짐을 알 수 있다. 따라서 각 사용자의 웹 페이지에 대한 데이터를 주성분으로 선형 변환하고 3개의 주성분에 의한 3차원 그래프를 그려보면 (그림 6)에서 보여지는 것처럼 전체적으로 3개의 군집이 형성됨을 시각적으로 관찰할 수 있다.



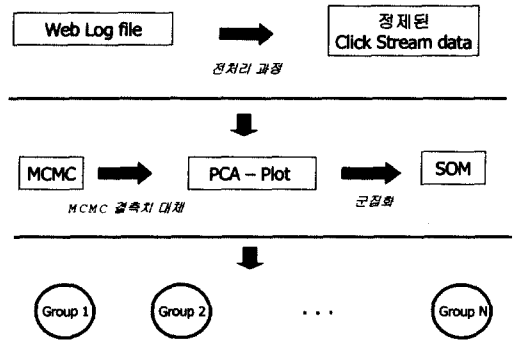
(그림 6) 주성분에 의한 군집수 결정

이 결과를 이용하여 SOM의 차원을 결정하게 된다. 본 논문에서는 (주성분의 개수×주성분의 개수)를 형상 지도의 차원으로 사용하였다. 여러 가능한 방법 중에서 이것이 군집의 결과가 가장 좋게 나왔다. 따라서 앞으로 수행될 다른 데이터의 군집화에도 이 방법을 사용하면 군집내의 동질성이 높은 군집 결과를 얻을 수 있을 것으로 기대된다. <표 4>는 제안하는 통계적 자율 신경망을 통하여 얻어진 군집화 결과이다.

〈표 4〉 SOM에 의한 군집화 결과

군 집	사용자 수	분 산
Group 1	6913	0.98
Group 2	4339	1.21
Group 3	1857	1.09

<표 4>에서 분산은 각 군집 별 269개의 웹 페이지에 대하여 각 페이지 마다의 머문 시간에 대한 분산의 평균값이다. 결론적으로 (그림 7)에서와 같이 희소한 웹 로그 데이터로부터 MCMC 결측치 대체 기법을 이용하여 희소성을 제거하고, 사전 군집수에 대한 정확한 정보를 주성분 분석에 의해 얻고 빠른 군집화 도구인 SOM을 이용하여 사용자를 군집화하였다.



(그림 7) 사용자 군집화

이러한 군집 결과를 이용하여 서로 유사한 집단별로 차별화된 웹 서비스를 수행하는 추천 시스템이나 사용자 모델링 등에 본 논문에서 제안하는 방법을 적용할수 있을 것이다. 또한 본 논문에서는 MCMC를 포함한 베이지안 추론과 주성분 분석의 학습 시간을 단축하기 위하여 전체 데이터를 사용하지 않고 Bootstrapping에 의한 재 표본(resampling) 방법을 이용하여 모형의 모수를 추정한 후 전체 데이터에 대한 학습을 수행하였다.

5. 결론 및 향후 과제

본 논문에서는 현재 웹 로그 데이터가 가지고 있는 가장 큰 어려움 중의 하나인 희소성 문제를 해결하기 위하여 통계 물리학에서 사용되는 MCMC 기법을 변형한 Hybrid MCMC 방법을 제안하였다. KDD Cup 데이터에 의한 실험을 통하여 기존의 웹 마이닝에서 결측치 대체 기법으로 사용되고 있는 회귀모형이나 Support Vector Machine 그리고 기존의 MCMC 방법 등과 비교한 결과 이들보다 정확한 예측력을 보이고 있음을 알 수 있었다. 또한 MCMC 기반의 베이지안 추론에서의 문제점인 모형에 대한 수렴을 위한 반복 학습 시간을 단축하기 위하여 Bootstrapping 기법을 적용하였다. 희소성을 제거한 웹 로그 데이터로부터 추천 시스템, 사용자 모델링 등에 적용할 사용자 군집화를 위하여 주성분 산점도 기반의 SOM을 적용하였다. 이 방법은 기존의 SOM의 시행 착오적인 형상 지도의 차원 결정 문제에 대한 해결 방안의 하나로써 주성분 분석을 사용한 것이다. 향후에 MCMC의 베이지안 추론에 있어서 좀더 빠른 수렴 알고리즘을 개발 적용하여, 정확한 예측력을 유지하면서도 현재의 Support Vector Machine 등의 방법 보다도 빠른 학습 시간을 갖는 모형에 대한 연구가 필요하다.

참 고 문 헌

[1] Sonny Han Seng Chee, "RecTree : A Linear Collaborative Filtering Algorithm," M. Sc. thesis, Dept. of Computer Science, Univ. Of Toronto, 1992.

[2] C. Guilfoyle, "Ventors of agent technology," in Proc. UNICOM Seminar Intell. Agents and Their Business Applicat., London, U. K., pp.135-142, 1995.

[3] J. Han, M. Kamber, "Data Mining : Concepts and Techniques," Morgan Kaufmann Publishers, 123-124 , 2001.

[4] W. J. Kennedy, Jr James E. Gentle, "Statistical Computing," Marcel Dekker, INC., 1980.

[5] T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, 43, pp.59-69, 1982.

[6] T. Kohonen, "Self-Organizing and Associative Memory," Springer, 1984.

[7] T. Kohonen, "Self Organizing Maps," Springer, 1997.

[8] T. M. Mitchell, "Machine Learning," McGraw-Hill, 1997.

[9] M. E. J. Newman, G. T. Barkema, "Monte Carlo Methods in Statistical Physics," Clarendon Press, 1999.

[10] S. M. Ross, "Introductory Statistics," McGrawHill, 1996.

[11] D. B. Rubin, "Multiple Imputation for Nonresponse in Surveys," John Wiley & Sons, Inc., 1987.

[12] B. M. Sarwar , G. Karypis, J. A. Konstan, J. Riedl , "Application of Dimensionality Reduction in Recommender System-A Case Study," WebKDD, Web Mining for E-Commerce Workshop, 2000.

[13] B. M. Sarwar, "Sparsity, Scalability, and Distribution in Recommender Systems," Ph. D. Thesis, Computer Science Dept., Univ. of Minnesota, 2001.

[14] J. L. Schafer, "Analysis of Incomplete Multivariate Data," Chapman and Hall, 1997.

[15] V. N. Vapnik, "Statistical Learning Theory," John Wiley & Sons Inc., 1998.

[16] http://www.ecn.purdue.edu/KDDCUP/.

전 성 해

e-mail : shjun@ailab.sogang.ac.kr

1993년 인하대학교 통계학과(학사)

1996년 인하대학교 대학원 통계학과 (이학석사)

2001년 인하대학교 대학원 통계학과 (이학박사)

2001년~현재 서강대학교 대학원 컴퓨터학과 박사과정

2003년~현재 청주대학교 통계학과 전임강사

관심분야 : 데이터마이닝, 기계학습, 인지과학

오 경 환

e-mail : kwoh@ccs.sogang.ac.kr

1978년 서강대학교 수학과(학사)

1985년 Florida State University Computer Science(공학석사)

1988년 Florida State University Computer Science(공학박사)

1989년~현재 서강대학교 컴퓨터학과 교수

관심분야 : 퍼지로직, 인공지능, 다중에이전트