Reversible Jump MCMC 와 베이지안망 학습에 의한 데이터마이닝

하 선 영[1] , 장 병 탁[1,2]

[1] 서울대학교 인지과학 협동과정

[2] 서울대학교 컴퓨터공학부

{syha, btzhang}@scai.snu.ac.kr

Data Mining Using Reversible Jump MCMC and Bayesian Network Learning

Sun-Young Ha[1], Byoung-Tak Zhang[1,2]

[1] Interdisciplinary Program in Cognitive Science

[2] School of Computer Engineering and Science

Seoul National University

요 약

데이터마이닝 문제는 데이터를 그 속성들에 따라 분류하여 예측하는 것뿐만 아니라 분류된 속성들간의 연관성에 대해 잘 설명할 수 있어야 한다. 일반적으로 변수들간의 연관성을 잘 설명할 수 있으면서도 높은 예측력을 가지는 방법으로는 베이지안 네트웍 분류자(Bayesian network classifier) 가 있다. 그러나 이것은 데이터 마이닝과 같은 대용량 데이터에서는 성능이 떨어지는 단점이 있다. 이에 이 논문에서는 최근 RBF 신경망의 입력변수 선정문제에 성공적으로 적용된 Reversible Jump Markov Chain Monte Carlo 방법을.이용하여 최적의 입력변수들만을 선택하여 베이지안 네트웍을 학습하는 Selective BN Augmented Naïve-Bayes Classifier 를 새로운 방안으로 제안하고 이를 실제 데이터마이닝 문제에 적용한 결과를 제시한다.

1. 서론

데이터마이닝(datamining)이란 대규모의 데이터베이스로부터 기존의 기법으로는 밝혀지지 않았던 해석가능한 유용한 정보를 추출해내는 과정을 말한다[1][2]. 여기서 해석가능하다는 것은 데이터분석의 결과가 통계적 용어나 기계학습 알고리즘에 익숙하지 못한 사람들도 쉽게 이해할 수 있어야 한다는 것을 의미한다.이런 이유로 데이터 마이닝에서 요구하는 분석의 결과는 데이터들간의 상관관계나 데이터를 비슷한 유형들의 집합으로 분류하는데서 끝나지 않는다. 예를 들어 자사의 인터넷 쇼핑몰의 성과를 알고자 하는 기업은 데이터마이닝 시스템이 자신의 홈페이지를 방문한 고객이 물건을 살 것인지 아닌지를 예측하고 분류할 수 있기만을 바라지 않는다. 그들은 왜 어떤 이유로 그런 사람들이 자신의 물건을 구입하는지에 대한 설명을 듣기 원한다.

이런 연구에 대표적인 시도가 바로 데이터마이닝에 베이지안 네트웍(Bayesian networks)을 도입한 것이다[3]. 베이지안 네트웍은 조건부 독립성을 나타내는 Directed Acyclic Graph(DAG)를 사용하여 많은 변수들간의 다양한 확률분포를 비교적 축약된 형태로 표현하기 때문에 변수들간의 상관관계를 쉽게 이해하고자 할 때 유용하게 쓰인다[4]. 데이터마이닝에 베이지안 네트웍을 적용할 때 문제가 되는 것 중에 하나는 일반적으로 데이터마이닝의 데이터들의 차원이 아주 크고 실제로 관심이 있는 속성들과 연관이 없는 정보들이 많다는 것이다. 이에 데이터의 차원 축소는 필수적이라 하겠다. 이 논문에서는 표본 기반의 입력변수집합을 선택하는 알고리즘을 베이지안 네트웍과 결합하여 능동적으로 학습을 하는 Selective Augmented Naïve-Bayes Classifier 를 제안하고 실험적 결과를 통해 이 방법이 데이터마이닝에 유용함을 보이고자 한다.

2. Selective Augmented Naïve-Bayes classifier

2.1 Bayesian Network Classifier

베이지안 네트웍 분류자(Bayesian Network Classifier)는 베이지안 네트웍을 분류문제에 적용한 것으로 목표값을 나타내는 노드가 모든 노드들의 부모노드가 되는 형태이다. 종류에는 Naïve-Bayes, Tree Augmented Naïve-Baye Classifier(TAN)[11], BN Augmented Naïve-Bayes Classifier(BAN)[12] 등이 있다. 특히[12]에서 제시된 BN Augmented Naïve-Bayes Classifier(BAN)는 베이지안 네트웍과 같은 구조를 허락하면서도 단지 $O(N^2)$의 계산복잡도를 가진다. 이에 여기서는 BAN 을 Classifier 로 이용하고자 한다.

2.2 BN Augmented Naïve-Bayes Classifier(BAN)의 학습

[12]에서는 BAN 의 학습 알고리즘으로 CBL 을 제시하고 있다. CBL algorithm 은 각 입력변수들간의 상호 연관관계를 검사하는 mutual information test(이것은 $O(N^2)$ 의 복잡도를 가진다)를 이용한 알고리즘으로 Drafting, Thickening 그리고 Thining 의 세단계로 이루어져 있다. Drafting 단계에서는 두 변수간의 mutual information 을 이용하여[1] 기초적인 네트웍의 초안을 그리게 된다. 이 때 그려지는 네트웍은 트리나 폴리트리의 모양을 형성하게 된다. 이 알고리즘에서 사용된 mutual information 은 식(1)

[1] Chow-Liu algorithm [13]에 의하면 두 변수의 conditional mutual information 의 값 $I(X_i, X_j \mid C)$ 이 일정 임계값보다 작으면 두 변수 X_i, X_j 를 C 에 의해 d-separation 이 된다. 이것은 두 변수가 조건부 독립임을 의미한다.

과 같은 conditional mutual information 이며 자세한 내용은 [12][13]에 나와있다.

$$I(X_i,X_j \mid C) = \sum_{x_i,x_j} P(x_i,x_j,c)\frac{P(x_i,x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)} \quad (1)$$

Thickening 이란 위에서 얻어진 초안에서 연결되지 않은 변수들간에 edge 를 추가하고, Thining 은 더해진 edge 들의 d-separate 여부를 판단하여 추가된 edge 를 제거 또는 보존하는 과정이다. 이 알고리즘은 계산이 빠르고 간편한 장점을 가지고 있으나 missing value 를 다룰 수 없고 데이터의 양이 충분히 많아야만 성립한다는 제약조건을 가지고 있다. 또한 목적값(target value)과 관련이 없는 속성이 많이 존재할 때는 성능이 떨어지는 단점이 있다.

2.3 Reversible Jump MCMC 를 이용한 입력변수선택

MCMC 와 같은 표본기반(sampling based)방법은 그동안 고정된 차원에서만 사용이 가능했기 때문에 model selection 이나 feature selection 과 같이 차원이 다른 공간들 사이를 이동해야 하는 과제에는 쓰일 수가 없었다. 그러나 [10]에 의해 Reversible Jump MCMC 가 소개됨으로써 이런 문제를 해결할 수 있게 되었다. Reversible Jump MCMC 는 기본적으로 Metropolis-Hastings 알고리즘에 기반한 것으로 서로 다른 입력변수집합들 사이를 Markov Chain 이 교대로 방문을 하면서 sampling 을 하게 된다. 이 때 각 차원을 이동하는 것은[10]에서 제안한 대로 생성이동(birth move), 소멸이동(death move), 높이이동(height move), 위치이동(position move)의 네 가지 이동에 의해서 이루어진다. 위 이동들은 각각의 수락확률 b_k, d_k, h_k, p_k 를 가지게 되는데 (k = 입력변수의 수, 즉 차원) $b_k + d_k + h_k + p_k = 1$ 을 만족하며 일반적으로 제안분포[2](proposal distribution) $q(u)$ 로부터 랜덤추출된 새로운 임의의 변수 u 가 b_k 와 같거나 작으면 생성이동, $b_k + d_k$ 보다 같거나 작으면 소멸이동, $b_k + d_k + p_k$ 보다 작거나 같으면 위치이동을 하게되고 그 이외의 경우에는 높이이동을 하게된다. 생성이동이란 새로운 입력변수를 랜덤하게 추출하여 현재의 입력변수집합의 차원(즉 입력변수의 수)을 증가시키는 것으로 이 때 수락확률[3] b_k 는 아래 식(2)와 같다.

$$b_k = \min\left\{1, \frac{p(y \mid A')}{p(y \mid A)} \frac{p(k+1)}{p(k)} \frac{q(u')}{q(u)}\right\} \quad (2)$$

여기서 A 는 현재입력집합을 의미하는 것으로 주어진 차원 k 안에서의 입력변수집합의 가능한 쌍들의 모임인 h 와 차원 k 를 정해진 스텝으로 나눈 위치를 나타내는 s 로 이루어진 공간에 위치한다. 소멸이동은 현재의 입력변수 집합에서 임의의 입력변수를 삭제하여 차원을 감소시키는 과정으로 이 때 거절확률은 위의 식(2)의 역인 $\frac{1}{b_k}$ 이다. 높이이동은 h 를 이동시키는 것으로 식(3)의 확률을 가지고 이동하며

$$h_k = \min\left[1, \frac{p(y \mid A')}{p(y \mid A)} \times \frac{h'^{\alpha}}{h^{\alpha}} \times \exp\{-\beta(h' - h)\}\right] \quad (3)$$

위치이동은 s 의 값을 변화시켜주는 것으로 식(4)의 확률로 이동한다.

$$p_k = \min\left[1, \frac{p(y \mid A')}{p(y \mid A)} \times \frac{(s_{j+1} - s'_j)(s'_j - s_{j-1})}{(s_{j+1} - s_j)(s_j - s_{j-1})}\right] \quad (4)$$

[2] 일반적으로 균일분포(uniform distribution)를 사용한다.

[3] 원래 식에서는 Jacobian 으로 나누어주게 되나 지금의 입력변수 선택 문제에서는 이 값이 1 이된다.

2.4 Selective BAN

이 논문에서는 위에서 언급된 두 방법을 결합하여 대용량 데이터를 효과적으로 다룰 수 있는 새로운 학습방법으로서 Selective BAN 을 제안한다. 이것은 다음과 같은 두단계로 이루어진다.

2.4.1 고정된 차원에서의 BAN 학습

Reversible Jump MCMC 에 의해 선택된 입력값에 대해 2.2 절에서 제시된 학습 알고리즘을 이용하여 베이지안 네트웍을 학습하게 된다. 학습이 끝나면 입력 값에 대한 최적의 베이지안 네트웍을 형성하게 되고 이 결과 $p(y \mid A)$ 의 값을 구할 수 있게 되며 이 값은 Reversible Jump MCMC 의 탐색에 다시 사용된다.

2.4.2 Reversible Jump MCMC 를 이용한 입력변수 집합 탐색

입력 집합의 크기에 대한 확률 $p(k)$ 는 [9]에서 제안된바와 같이 식(3)과 같은 truncated Poisson 분포를 사용한다.

$$p(k) = 1/\frac{k!}{(k-k_{\max})!k_{\max}!} c\frac{\lambda^k}{k!} \quad (5)$$

여기서 $k_{\max}$ 는 입력값의 최대수를 의미한다. 이 식을 이용하여 2.3 절에서 설명되어진 대로 입력집합을 찾아나가며, 목적값과 연관된 최적의 입력변수집합의 선택은 얼마나 자주 특정 집합이 방문되어졌는가로 계산된다.

3. 실험

3.1 데이터

실험에 사용한 데이터는 KDD2000 Competition 데이터 중 Question 3 에 대한 것이다. Question 3 는 Legcare 전문회사의 홈페이지를 방문한 사용자들의 기록(회원등록기록, 방문한 페이지 등)을 바탕으로 평균적으로 12 달러 이상을 소비하는 방문자(heavy spender)에 대한 특성을 설명하는 것이다. 이 설명은 마케팅 전문가들이 실제로 이해하고 받아들일 수 있는 결과이어야 한다. 이 데이터의 전체 속성은 465 개 이며 이산과 연속 변수들이 섞여있다. 전체 데이터 수는 1782 개이다. 이중 약 80% 인 1420 개를 학습 데이터로 나머지를 테스트 데이터로 사용하였다. 모든 missing value 는 평균값으로 대체되었다.

3.2 실험방법 및 성능 측정 방법

Selective BAN 은 λ =3, $k_{\max}$ =30, α =1, β =200, c =0.25 로 고정한 후, 초기 4000 개의 샘플을 burn-in 한 후 40000 번을 수행하였다. 이 후 학습된 Selective BAN 에 대해 테스트 데이터를 이용하여 분류에 대한 예측도를 측정한 후 이를 Decision Tree, Neural Network, Bayesian Neural Network with ARD 와 비교하였다.

변수간 관계를 찾아내는 능력에 대한 성능평가는 KDD 2000 Competition 의 채점 기준에 따라 행하여졌으며 여기서 얻어진 점수를 Competition 우승자 및 다른 참가자들의 점수와 비교하였다.

4. 결과 및 토의

4.1 실험 결과

Selective BAN 은 표 1 에서 보이는 것과 같이 예측도에 있어서 비교적 좋은 결과를 나타내었다. 표에서 ARD 를 사용한 Bayesian Neural Network 이 제일 좋은 결과를 나타내었는데 ARD 를 사용하면 학습과 동시에 feature subset selection 을 행할 수 있기 때문이며 이는 KDD 데이터처럼 목적값과 상관없는 데이터가 많을 경우 더 효과적이다. 그러나 이 방법은 다른 방법들과 마찬가지로 입력변수들간의 상관관계에 대해서는 설명을 할 수가 없다.

표 1. 12 달러 이상을 소비한 소비자에 대한 예측도 (단위 %)

Selective BAN	Decision Tree	Neural Network	Bayesian Neural Network with ARD
84.20	82.50	82.08	84.33

표 2 는 다른 입력변수 선택 알고리즘에 의해 구해진 입력 변수들과 Selective BAN 에 의해 구해진 입력변수집합을 비교하기 위한 것이다. 칸 왼쪽은 KDD 채점기준에 포함된 변수들의 목록이다. 각 칸마다 이 변수들이 선택된 입력변수집합에 포함되어 있는지의 여부를 체크했다. 여기서는 KDD 채점기준에 포함된 입력변수를 더 많이 찾아낸 알고리즘이 더 우수하다고 보았다.

표 2. 입력변수선정에 대한 비교

	Selective BAN	Decision Tree	Bayesian Neural Network with ARD
Friend Promotion	▯	▯	▯
Time effect	▯	×	×
My coupon	▯	×	▯
Send e-mail	▯	▯	▯
Buy for others	▯	▯	▯
Product view	▯	▯	▯
Work dress	▯	×	×
Geography	▯	×	×
Causual socks	▯	×	×
Leg care	×	×	×
Year of date	×	×	×
Pantyhose	×	×	×

그림 2 는 Selective BAN 에 의해 구성된 베이지안 네트웍이다. 여기서 학습된 BAN 을 통해 찾아낸 사실은 아래와 같다

1. Friend 프로모션을 통해 이 웹사이트에 들어온 사람은 My coupon 과 Family 프로모션에도 영향을 받으며 주로 실용적인 옷을 입는 사람들로 heavy spender 일 확률이 낮다.
2. 위치상으로 Northeast U.S.에 사는 사람일수록 수입이 높으며 주로 다른 사람들을 위해 물건을 구입한다. 이들은 heavy spender 이다.
3. Product view 중 도나 카렌과 같은 상품을 보는 사람들은 주로 Northeast U.S.에 살며 제품 설명을 받아보고 싶다는 e-mail 을 보내며 heavy spender 이다.

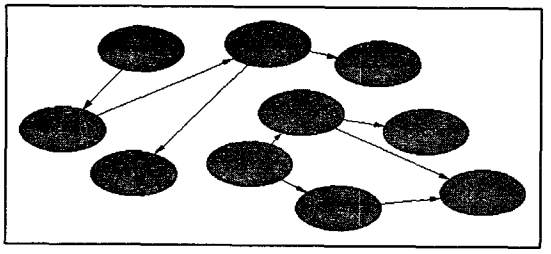

그림 2. Selective BAN 에 의해 학습된 베이지안 네트웍 (목적노드는 제외하였음)

4.2 토의

위의 결과는 KDD 의 채점기준에 의하면 30 점에 해당하는 것으로 다른 참가자들과의 점수를 비교하여 중간정도라고 할 수 있다. 이는 각 변수들 사이의 은닉 요소를 전혀 고려하지 않은 탓이며 입력변수의 최대값을 30 으로 제한한 까닭도 있는 듯하다. 그러나 찾아진 입력변수의 내용에서 보면 다른 알고리즘에 비해 뛰어남을 알 수 있다. 이는 Selective BAN 이 목적값과 상관있는 변수만을 찾아서 학습할 수 있다는 것을 의미한다.

이 논문에서 제시된 Selective BAN 은 아직 견고하게 완성된 알고리즘이 아니다. 이런 이유로 많은 한계점을 가지고 있다. 첫째는 missing value 를 다룰 수 없다는 것이다. 이것은 데이터 마이닝 데이터와 같이 에러가 많은 데이터를 다룰 때는 심각한 문제가 아닐 수 없다. 두번째는 위계적 구조를 다룰 수 없다는 것이다. 위계적 구조를 가지면 입력변수들간의 숨은 원인을 찾아낼 수 있으며 좀 더 효율적으로 학습할 수 있는 장점을 가지게 된다. 위의 두가지 한계점은 지금 연구중에 있으며 [6]에서 이용된 방법 등이 적용하기에 적당하리라 생각한다.

감사의 글

본 연구는 정통부 대학기초연구(과제번호 CI-98-006800),학술진흥재단 자유공모과제(과제번호 1999-001-E01025), 첨단정보기술연구소(AITRC)에 의하여 일부 지원되었음.

참고 문헌

[1] Berry, M. and Linoff, G., *Data Mining Techniques*, Wiley, New York, 1997.

[2] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees*, Chapman & Hall, New York, 1984.

[3] Heckerman, D., "Bayesian networks for knowledge discovery," In *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[4] Heckerman, D., Meek, C., and Cooper, G., "A Bayesian Approach to Causal Discovery," *Technical Report MSR-TR-97-05*, Microsoft Research, February, 1997.

[5] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford, 1995.

[6] Koller, D. and Sahami, M," Hierarchically classifying documents using very few words," *Proceedings of the 14th International Conference on Machine Learning (ICML)*, Morgan Kaufmann, pages 170-178, 1997.

[7] D. Koller and M. Sahami, "Toward optimal feature selection," *Proceedings of the 13th International Conference on Machine Learning (ICML)*, Morgan Kaufmann, pages 284-292, 1996.

[8] Yang, J. and Honavar, V., "Feature subset selection using a genetic algorithm," In: *Feature Extraction, Construction and Selection - A Data Mining Perspective.* Motoda and Liu (eds.) Kluwer Academic Publishers, Chapter 8:117-136, 1998.

[9] Sykacek, P., "On input selection with reversible jump Markov chain Monte Carlo sampling," *In Advances in Neural Information Processing Systems 12* , S.A. Solla, T.K.Leen and K.-R. Mueller (eds.), MIT Press, pages 638-644, 2000.

[10] Green, P., "Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrica* 82, 711-732. 1995.

[11] Friedman, N., Geiger, D., and Goldszmidt, M., "Bayesian Network Classifiers," *Machine Learning 29*:131-163, 1997.

[12] Cheng, J. and Greiner, R., "Comparing Bayesian network classifiers," *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligenc,"* 1999.

[13] Cheng, J., Bell, DA., and Liu, W., "Learning belief networks from data: an information theory based approach," *Proceedings of the Sixth ACM International Conference on Information and Knowledge Management*, 1997.