Journal of the Korean
Data & Information Science Society
2001, Vol. 12, No. 2, pp. 11 ~ 25

베이지안 신경망을 이용한 분류분석 [1]

황진수 [2] · 최성용 [3] · 전홍석 [4]

요 약

자료들 사이에 존재하는 관계, 패턴, 규칙등을 찾아내서 모형화 하는 통계적인 분류기법은 여러가지가 있다. 그러나 우리가 얻게 되는 지식은 어떤 일련의 분류규칙에 의해서가 아닌 관찰과 학습을 통한 훈련으로부터 얻게 된다. 본 베이지안 학습은 모든 형태의 불확실성을 표현하는 확률로써 우리의 믿음의 정도를 표현하는 것으로 해석될 수 있으며, 확실한 결과가 알려짐에 따라 확률이론 법칙을 사용하여 이러한 확률들을 갱신한다. 또한 신경망 모형은 이미 알고 있는 속성들에 근거하여 아직 알지 못하는 집단이나 특질들을 예측하게 해준다. 본 논문에서는 이러한 두 가지 방법을 결합한 베이지안 신경망과 기존의 CHAID, CART, QUEST 분류 알고리즘에 있어서 각각 오분류율을 비교연구하였다.

주제어: ARD, 분류, 베이지안 신경망, 하이브리드 몬테칼로

제 1 절 서 론

관찰로부터 학습을 하는 능력은 모든 지식의 근원이다. 우리는 냄새로 와인의 품종과 생산 년도를 알아내고, 손으로 쓴 글자나 숫자를 인식하며, 크기와 외형으로부터 사물의 무게를 예측하는 등 경험적 자료로부터 보이지 않는 특성들을 얻어낸다. 학습의 이론과 방법론들은 다양한 분야에서 관심의 대상이 되고 있다. 이러한 학습과 다른 형태의 추론은 이론상 단순한 확률법칙에 의해서 수행될 수 있다. 베이지안 학습은 모든 미지의 특성치에 대하여 여러 가능성에 대한 우리의 믿음의 정도를 확률분포의 형태로 표현하며, 신경망은 이미 알고 있는 속성들에 근거하여 아직 알지 못하는 집단이나 특질들을 예측하게 해준다. 최근 신경망은 많은 분야에서 성공적으로 응용되면서 실제적인 방법으로 출현하였다. 이러한 신경망 활용의 대부분은 패턴인식과 관련되어 있고 다층인식 네트워크(multi-layer perceptron network)와 RBF 네트워크(radial basis function network)와 같은 전방향 네트워

[1]본연구는 2000년도 인하대학교의 지원에 의하여 연구되었음(INHA-21072)
[2](402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 부교수
[3](402-751) 인천광역시 남구 용현동 253, 인하대학교 전자계산공학과 박사과정
[4](402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 교수

크(feed forward network)를 사용한다. 이러한 신경망의 성공적인 활용에는 보다 이론적인 접근이 필요하다는 공통된 인식이 생기게 되었다. 역사적으로 볼 때 신경망의 많은 개념들은 생물학적 네트워크에 의해 생성되었지만 통계학적 패턴인식의 관점은 더 직접적이고 이론적인 방향을 제시해 준다.

본 논문에서는 이러한 두 가지 방법을 합한 베이지안 신경망을 사용한 분류모형과 CHAID, CART, QUEST의 다른 알고리즘을 사용한 분류모형에서의 오분류율을 비교 연구하였다.

제 2 절 이론적 배경

2.1 베이지안 신경망

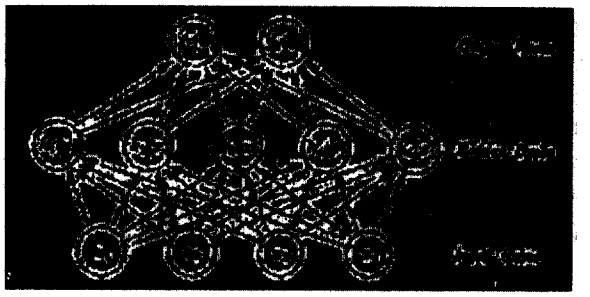

그림 1 : 다층인식 네트워크

여기서 논의될 신경망은 역전파(*back-propagation*) 혹은 전방향(*feed forward*) 네트워크라고 알려진 다층인식네트워크(Rumelhart, Hinton, and Williams 1986a, 1986b)이다. 이 네트워크는 실수 입력값들의 집합 x_i와 몇 개의 은닉 단위들의 층을 사용하여 그들로부터 하나 혹은 그 이상의 출력값 $f_k(x)$들을 계산한다. 위의 그림에서와 같이 1개의 은닉 층을 가진 전형적인 네트워크에서 출력값들은 다음과 같이 계산된다.

$$f_k(x) = b_k + \sum_j v_{jk} h_j(x) \tag{2.1}$$

$$h_j(x) = \tanh(a_j + \sum_i u_{ij} x_i) \tag{2.2}$$

여기서 u_{ij}는 i번째 입력단위로부터 j번째 은닉단위로 연결되는 가중치이고 이와 유사하게, v_{jk}는 j번째 은닉단위로부터 k번째 출력단위로 연결되는 가중치이다. a_j 와 b_k는 은닉단위와 출력단위의 편의(bias)들이다. 이들 가중치들과 편의들은 네트워크의 모수들이다. 실수값을 반응변수로 갖는 회귀모형에서, 입력 x가 주어졌을 때 반응변수 y_k의 조건부 분포는 평균이 $f_k(x)$이고 표준편차 σ_k를 갖는 정규분포로 가정 한다. 각각의 결과 값은 입력이 주어졌을 때 아래와 같이 항상 독립적으로 얻어진다.

$$p(y|x) = \prod_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-(f_k(x) - y_k)^2 / 2{\sigma_k}^2) \tag{2.3}$$

K개의 가능한 클래스(class)중 하나를 가리키는 이산형 값을 반응변수로 갖는 분류모형에서, 반응값 y는 Softmax 모형(Birdle 1989)에서와 같이 K개의 출력단위를 가진 네트워크를 사용하여 여러 클래스의 조건부 확률을 정의 하는데 사용될 수 있다.

$$p(y=k|x) = \frac{\exp(f_k(x))}{\sum_{k'} \exp(f_{k'}(x))} \tag{2.4}$$

신경망에서 가중치와 편의는 입력값 $x^{(i)}$와 그에 관련된 반응값 $y^{(i)}$로 구성된 훈련자료(training data) $(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})$에 기초하여 얻어진다. 신경망 학습의 기본과정은 네트워크의 가중치와 편의를 훈련자료에서 결과값과 반응값 간의 차이 제곱인 오차를 최소화 하도록 조정하는 것이다. 또한 신경망 학습에서의 베이지안 접근에서 목표가 되는 것은 새로운 입력이 주어지고, 훈련자료에서 입력과 반응값이 주어졌을 때 새로운 검정자료(test data)에서의 반응값에 대한 예측분포를 찾는 것이다. 새로운 입력값 $x^{(n+1)}$이 주어졌을 때 출력값 $y^{(n+1)}$에 대한 예측분포는 베이지안 구조에서 다음과 같이 표현된다.

$$\begin{aligned} &p(y^{(n+1)}|x^{(n+1)}, (x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})) \\ &\quad = \int p(y^{(n+1)}|x^{(n+1)}, \theta) p(\theta|(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)}))d\theta \end{aligned} \tag{2.5}$$

여기서 θ는 네트워크에서의 모든 가중치와 편의를 나타내는 네트워크 모수이다. 입력값에 대한 분포가 모형화 되지 않기 때문에 θ의 가능도함수는 다음과 같다.

$$L(\theta|(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)}, \theta) \tag{2.6}$$

입력, $x^{(i)}$와 네트워크 모수가 주어졌을 때 반응값, $y^{(i)}$의 분포는 회귀모형과 softmax분류모형에서 (2.3), (2.4)로 주어진 것처럼 사용된 네트워크의 모형형태에 의해 정의된다. 제곱오차 손실함수를 사용하여 $y^{(n+1)}$을 예측할 때 가장 좋은 방법은 예측분포의 평균을 추정하는 것이다. 회귀모형에서는 다음과 같은 추정치를 만들어 낸다.

$$\hat{y}_k^{(n+1)} = \int f_k(x^{(n+1)}, \theta) p(\theta|(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)}))d\theta \tag{2.7}$$

2.2 하이브리드 몬테칼로 알고리즘

하이브리드 몬테칼로 알고리즘 (Duane, Kennedy, Pendleton, and Roweth ,1987) 은 메트로폴리스 알고리즘과 역학적 모의실험에 기초한 표본추출방법을 결합한 것이다. 이러한 하이브리드 몬테칼로 알고리즘의 결과는 어떤 특정한 분포로부터 발생시키는 표본이 되며 그러한 분포들에 대한 여러 다른 함수들의 기대값의 추정치로써 사용될 수 있다. 즉 베이지안 학습에서는 훈련자료가 주어진 사후확률분포로부터 표본을 추출하는 것이며 (2.7)에서와 같이 검정자료에서 예측을 하는데 필요한 기대값을 추정하는 것이다.

하이브리드 몬테칼로 알고리즘은 확률함수의 미분값이 계산될 수 있는 정준분포로부터 표본추출하는 것으로 표현되며 이 때의 정준분포는 에너지 함수로써 정의 된다. 만약 N개

의 요소 q_i로 구성된 위치변수(position variable) q의 어떤 분포로부터 표본추출한다고 가정할 때 정준분포는 다음과 같이 정의된다. (본 논문에서 q_i들은 네트워크의 모수가 된다.)

$$P(q) \propto \exp(-E(q)), \text{여기서 } E(q)\text{는 에너지 함수이다.} \tag{2.8}$$

윗 식의 에너지함수 $E(q)$는 다음과 같이 다시 표현된다.

$$E(q) = -\log P(q) - \log Z, \text{여기서 } Z\text{는 정규화 상수이다.} \tag{2.9}$$

역학적방법을 사용하기 위해 N개의 요소 p_i로 구성되며 q_i의 구성요소와 일대일로 대응되는 운동량 변수(momentun variable) p가 필요하다. q와 p로 구성된 상태공간에서의 정준분포는 다음과 같이 정의된다.

$$P(q,p) \propto \exp(-H(q,p)) \tag{2.10}$$

이 때 해밀토니안 함수 $H(q,p)$는 $E(q)+K(p)$로 구성되며 전체에너지를 나타낸다. $K(p)$는 운동량에 따른 운동 에너지로써 다음과 같이 계산된다.

$$K(p) = \sum_{i=1}^{n} \frac{p_i^2}{2m_i} \tag{2.11}$$

여기서 m_i 는 각 구성요소들에 관련된 질량들을 말하며 예측에 대한 정확성을 높이기 위해서는 이들 질량값의 조정이 필요하다. (2.10)의 분포에서 q와 p는 독립이며 q의 주변분포는 (2.8)과 동일하다. 즉, q와 p에 대해 정준분포로 수렴하는 마코프 체인을 정의한 후, p의 값을 무시함으로써 함수 q의 기대값의 추정등을 할 수 있다. (2.10)에서의 고정된 전체 에너지로부터 표본추출하는 방법은 가상의 시간 τ에 따라 다음 식에 의하여 상태가 변하게 되는 해밀토니안 역학에 의해서 얻어진다.

$$\frac{dq_i}{d\tau} = +\frac{\partial H}{\partial p_i} = \frac{p_i}{m_i}$$

$$\frac{dp_i}{d\tau} = -\frac{\partial H}{\partial q_i} = -\frac{\partial E}{\partial q_i}$$

이러한 연속적 해밀토니안 역학을 이산형의 형태로 근사시키는 방법을 leapfrog 스텝이라 부르며 이러한 leapfrog 스텝은 시간 τ에서의 $\hat{q}$와 $\hat{p}$로부터 스텝크기 ϵ후인 $\tau+\epsilon$에서의 $\hat{q}(\tau+\epsilon)$와 $\hat{p}(\tau+\epsilon)$의 근사치에 해당되는 하나의 단계를 얻을 수 있다.

$$\hat{p}_i(\tau+\frac{\epsilon}{2}) = \hat{p}_i(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial q_i}(\hat{q}(\tau))$$

$$\hat{q}_i(\tau+\epsilon) = \hat{q}_i(\tau) + \epsilon\frac{\hat{p}_i(\tau+\frac{\epsilon}{2})}{m_i}$$

$$\hat{p}_i(\tau+\epsilon) = \hat{p}_i(\tau+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial E}{\partial q_i}(\hat{q}(\tau+\epsilon))$$

이러한 leapfrog 스텝이 진행된 후의 $H(q,p)$의 값은 연속인 헤밀토니안 시스템과는 달리 다른 값으로 변할 수가 있다. 이러한 이유로 위와 같은 방법에 의하여 얻어진 몬테칼로 추정치

는 체계적인 오차를 발생시키게 되지만 이러한 체계적인 오차는 메트로폴리스 알고리즘을 사용하여 제거될 수 있다. 즉, 이러한 역학에 의해 도달된 상태에서의 전체에너지 $H(q,p)$값의 변화에 기초하여 메트로폴리스 알고리즘을 사용하여 채택과 기각 여부를 결정한다. 만약, 역학의 이동이 정확하게 구현되었다면 $H(q,p)$에서의 변화는 항상 0이 되며 그때의 상태는 항상 채택하게 된다. 만약 역학의 이동이 정확하게 구현되지 못했다면 $H(q,p)$의 값은 변하게 될 것이고 그때의 이동은 일정한 확률로 기각되어 이전 상태로 되돌리게 된다. 이러한 기각은 부정확한 구현에 의해 발생된 체계적인 오차를 제거하는 역할을 한다.

2.3 하이퍼파라미터를 사용한 베이지안 신경망

베이지안 신경망은 입력으로부터 출력으로 계산되는 함수들 $f(x,\theta)$에 의해 정의된 네트워크 모수 θ로 표현되는 가중치와 편의들로 구성된다. 이들 네트워크 모수들의 사전확률은 하이퍼파라미터 γ의 값에 따라 정의된다. 즉 모수들의 사전확률분포는 $p(\theta|\gamma)$와 같이 표현되며, 하이퍼파라미터 자체의 사전확률은 $p(\gamma)$이다. 또한 서로 독립된 입력값들로 구성된 훈련자료 $(x^{(1)},y^{(1)}),\cdots,(x^{(n)},y^{(n)})$에서 하이퍼파라미터 $p(\gamma)$를 사용한 반응값의 조건부분포를 모형화 할 때의 조건부확률 또는 조건부분포는 $p(y|x,\theta,\gamma)$와 같이 표현된다. 이때의 최종목표는 훈련자료에서의 정보를 이용하여 새로운 입력 $x^{(n+1)}$이 주어졌을 때 새로운 검정자료에 대한 반응값 $y^{(n+1)}$을 예측하는 것이다. 이러한 예측은 훈련자료에서의 사전확률과 가능도값의 곱에 비례하는 θ와 γ의 사후확률분포에 기초하여 다음과 같이 얻어진다.

$$p(\theta,\gamma|(x^{(1)},y^{(1)}),\cdots,(x^{(n)},y^{(n)})) \propto p(\gamma)p(\theta|\gamma)\prod_{c=1}^{n} p(y^{(c)}|x^{(c)},\theta,\gamma) \tag{2.12}$$

이 때 예측값 $y^{(n+1)}$의 분포는 위의 사후확률분포에 대하여 다음과 같이 적분하여 얻을 수 있다.

$$\begin{aligned} &p(y^{(n+1)}|x^{(n+1)},(x^{(1)},y^{(1)}),\cdots,(x^{(n)},y^{(n)})) \\ &\quad = \int p(y^{(n+1)}|x^{(n+1)},\theta,\gamma)p(\theta,\gamma|(x^{(1)},y^{(1)}),\cdots,(x^{(n)},y^{(n)}))d\theta d\gamma \end{aligned} \tag{2.13}$$

회귀모형에서 기대제곱오차를 최소화시키는 것은 이러한 예측분포의 평균이다. 만약, 반응값의 조건부분포가 네트워크 출력값에 일치하는 평균을 가지도록 정의된다면 최적의 예측은 다음과 같다.

$$\hat{y}^{(n+1)} = \int f(x^{(n+1)},\theta)p(\theta,\gamma|(x^{(1)},y^{(1)}),\cdots,(x^{(n)},y^{(n)}))d\theta d\gamma \tag{2.14}$$

베이지안 신경망에서는 사후확률분포의 복잡성으로 이러한 적분이 쉽지 않기 때문에 Markov Chain Monte Carlo(이 후에는 MCMC로 부름)방법을 사용한다. MCMC방법에서는 사후확률분포에 관한 함수의 기대값의 형태를 취하는 위와 같은 적분은 사후확률분포로부터 표본추출한 값들의 평균값에 의해 근사 될 수 있다. 이러한 사후확률분포로부터 표본추출하려고 할 때 조건부분포로부터 표본추출하는 것이 사실상 불가능하기 때문에 단순한 깁스

샘플링 방법은 사용할 수 없고, 기본 형태의 메트로폴리스 알고리즘은 가능하지만 임의보행의 문제점이 발생할 수 있으므로, Stochastic dynamics 방법을 보완한 하이브리드 몬테칼로 방법을 사용한다.

2.4 하이퍼파라미터와 네트워크모수의 갱신

베이지안 신경망모형에는 네트워크 모수의 사전확률분포와 회귀모형에서 *noise level*을 정의하는 두 종류의 하이퍼파라미터들이 존재한다. 전자의 하이퍼파라미터들은 사전지식에 기초한 모형의 선택에 따른 특정한 그룹에서의 모든 모수들에 대한 표준편차를 조절한다. 구체적으로 특정한 그룹에서의 θ의 구성요소가 되는 모수들 $u_1, \cdots, u_k$은 서로 독립이고 평균이 0, 표준편차 σ_u인 정규분포를 따른다고 가정할 때, 이러한 표준편차는 $\tau_u = \sigma_u^{-2}$으로 정의되는 정도(*precision*)로 표현되며, 그 그룹에서의 모수들에 대한 분포는 다음과 같다.

$$p(u_1, \cdots, u_k|\tau_u) = (2\pi)^{-k/2}\tau_u^{k/2}\exp(-\tau_u \sum_i u_i^2/2) \tag{2.15}$$

이 때 정도 τ_u는 평균이 w_k이고 형상(shape) 모수는 α_u를 갖는 감마분포로 표현되며, 그때의 확률밀도함수는 다음과 같다. 여기서 τ_u는 하이퍼파라미터 γ의 구성요소이다.

$$p(\tau_u) = \frac{(\alpha_u/2w_u)^{\alpha_u/2}}{\Gamma(\alpha_u/2)}\tau_u^{\alpha_u/2-1}\exp(-\tau_u\alpha_u/2w_u) \tag{2.16}$$

$u_1, \cdots, u_k$의 값이 주어졌을 때 τ_u에 대한 사후분포는 (2.15)와 (2.16)의 곱에 비례하여 다음과 같이 감마분포로 표현된다.

$$\begin{aligned} & p(\tau_u|u_1, \cdots, u_k) \\ & \quad \propto \tau_u^{\alpha_u/2-1} exp(-\tau_u\alpha_u/2w_u)\tau_u^{k/2} exp(-\tau_u \sum_i u_i^2/2) \\ & \quad \propto \tau_u^{(\alpha_u+k)/2-1} exp(-\tau_u(\alpha_u/w_u + \sum_i u_i^2)/2)) \end{aligned} \tag{2.17}$$

윗식의 조건부분포에서 τ_u의 값은 다른 모수, 하이퍼파라미터, 반응값들과 독립으로 깁스 샘플링으로 갱신된다.

후자의 하이퍼파라미터들도 이와 마찬가지로 k번째 반응값에 관련된 모형 오차의 표준편차가 σ_k인 정규분포를 따른다고 가정할 때 이러한 표준편차는 $\tau_k = \sigma_k^{-2}$으로 정의되는 정도로 표현되며, 이 때 입력값 $x^{(1)}, \cdots, x^{(n)}$, 네트워크모수 θ, 모형 오차의 정도 τ_k가 주어졌을 때 반응값의 확률분포는 다음과 같이 표현된다.

$$\begin{aligned} & p(y_k^{(1)}, \cdots, y_k^{(n)}|x_k^{(1)}, \cdots, x_k^{(n)}, \theta, \tau_k) \\ & \quad = (2\pi)^{-n/2}\tau_k^{n/2} exp(-\tau_k \sum_c (y_k^{(c)} - f_k(x^{(c)}, \theta))^2/2) \end{aligned} \tag{2.18}$$

또한 모형의 정도 τ_k의 사전분포를 다음과 같이 평균 w와 형상모수 α를 갖는 감마분포로 표현하면 아래와 같다.

$$p(\tau_k) = \frac{(\alpha/2w)^{\alpha/2}}{\Gamma(\alpha/2)} \tau_k^{\alpha/2-1} exp(-\tau_k \alpha/2w) \tag{2.19}$$

또한, τ_k에 대한 사후분포도 다음과 같이 감마분포로 표현된다.

$$\begin{aligned} & p(\tau_k|(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)}), \theta) \\ & \quad \propto \tau_k^{(\alpha+n)/2-1} exp(-\tau_k(\alpha/w) + \sum_c (y_k^{(c)} - f_k(x^{(c)}, \theta))^2/2) \end{aligned} \tag{2.20}$$

사후확률분포 전체를 이동하게 되는 마코프체인은 이전에서와 같이 하이퍼파라미터들에 대한 깁스샘플링과 네트워크 모수들에 대한 하이브리드 몬테칼로 방법을 교대로 시행하여 얻을 수 있다. 이러한 하이브리드 몬테칼로 방법은 네트워크 모수 θ의 함수인 에너지로 정의된 사후확률분포로부터 표본추출하는 것으로서 하나의 하이브리드 몬테칼로 갱신 동안 하이퍼파리미터들은 고정되며, 이 때 (2.12)는 에너지의 용어로서 다음과 같이 표현된다.

$$E(\theta) = F(\gamma) - \log P(\theta|\gamma) - \sum_{c=1}^{n} \log P(y^{(c)}|x^{(c)}, \theta, \gamma) \tag{2.21}$$

$\exp(-E(\theta))$에 비례하는 이와 같은 에너지 함수의 정준분포는 하이퍼파라미터 γ가 주어졌을 때 θ에 대한 사후확률분포를 생성한다. 즉, 위의 에너지함수는 하이퍼파라미터들이 바뀜에 따라 변하게 된다. 이러한 사후확률분포 전체를 이동하게 되는 마코프체인은 각 상태들간의 종속을 낮게 유지하여, 추정된 몬테칼로 방법의 정확도를 높이도록 leapfrog 이산화에서의 단계의 크기 ϵ과 단계의 수 L을 정한다.

2.5 Automatic Relevance Determination(ARD) 모형

베이지안 신경망에서 고려해야 하는 또 다른 문제는 반응값(target)의 분포를 모형화 시킬 때 사용되는 입력 변수들의 갯수이다. 많은 문제에서 입력변수를 추가시킨다면 예측수행능력을 향상시킬 수 있을 것 같은 측정 가능한 속성들이 있을 수 있다. 하지만 작은 훈련자료에서 관련이 없는 입력 변수들이 실제로 관련이 있는 변수들보다도 더 관련이 있는 것처럼 영향을 줄 수 있게 되기 때문에 점점 더 많은 입력 변수들을 추가 시키는 것은 결국에는 그러한 예측수행능력을 낮추게 된다. 따라서 가장 관련이 있을 것이라고 생각되는 입력변수들만을 고려해야 할 것이다. 하지만 우리가 예측하려고 하는 모든 것에 대한 완벽한 이해 없이는 과연 어떠한 속성들이 실제로 관련이 있는가에 대한 확신이 있을 수 없다. 그러므로 관련성이 알려지지 않은 입력변수가 어느 정도 관련이 있는 가를 자동적으로 결정하도록 하는 것이 필요하게 되었다.

ARD모형은 많은 입력 변수들 중 어떤 변수가 반응값(target)과 관련이 있는지를 자동적으로 결정하도록 하기 위해 Mackay(1994a, 1995b)와 Neal(1994)에 의해 개발되었다. 그 방법은 각 입력 변수들과 연결되는 가중치가 그 입력에 관련된 하이퍼파라미터에 의해 조

절되는 분포를 만드는 것이다. 그럼으로써 각 입력의 관련성이 자료에 부합하는 이러한 하이퍼파라미터들의 값에 따라 자동적으로 결정되도록 하는 것이다. 많은 입력변수들을 포함한 문제에서 각 입력변수들에 관련된 하이퍼파라미터들에 대해 사전정보를 가지는 사전확률을 사용할 필요가 있다. 그러나 만약, 어떤 특정한 입력변수에 대해 작은 사전확률을 갖는 하이퍼파라미터가 사용된다면 그 때의 입력변수와 연결된 가중치들은 매우 작게 될 것이며, 비록 그 때의 입력변수가 관련이 있는 변수라고 하더라도 그 때의 가중치들의 값이 너무 작아서 많은 자료가 주어진다 해도 관련이 있는 것으로 여겨질 수 없기 때문이다. 그러므로 각 입력변수의 하이퍼파라미터들에 대해 서로 다른 사전정보를 갖는 사전확률을 적용시키기 보다는 전체 자료에서 입력변수들을 정규화하는 (즉, 전체 훈련자료에서 각 입력변수들의 평균을0, 표준편차를 1로 만드는) 방법 (Quinlan , 1993)에 의해 각 입력 변수들에서의 변화가 다른 입력 변수들과 동등한 영향력을 갖도록 변화시킨 후 그들 모두에 같은 사전확률을 사용하여 모든 입력변수들에 대한 공통된 상위단계 하이퍼파라미터를 지정하고, 그 상위단계 하이퍼파라미터에 의해 결정되는 사전확률의 평균으로 다소 구체화된 값으로 하위 단계 하이퍼파라미터들을 지정하는 2단계의 하이퍼파라미터를 사용한다.

제 3 절 모의실험결과 및 검토

모의 실험에 사용된 자료는 보통 분류분석자료로 사용되는 자료로서 인터넷등을 통하여 얻은 총 5종류의 자료들이다. 이러한 자료들에 대하여 ARD 모형을 이용한 베이지안 신경망과 분류 알고리즘인 CHAID, CART, QUEST의 방법들에서 각각의 오분류율을 비교연구하였다. 사용된 베이지안 신경망은 Neal의 프로그램인 Software for Flexible Bayesian Modeling 이며, 비교연구된 분류알고리즘인 CHAID, CART, QUEST의 방법들은 SPSS사의 AnswerTree 2.0을 사용하여 분석하였다. 이 때 각 자료의 시드(SEED)를 바꾸어 가면서 총 20회의 반복실험을 하였다.

3.1 3WAY DATA(총 자료수 1000개)

자료설명 - 4개의 입력변수 $(x1, x2, x3, x4)$를 가지며 모두 일양분포 $[0, 1]$에서 임의 추출한다. 이때 점 $(x1, x2)$과 점 $(0.4, 0.5)$사이의 거리가 0.35보다 작으면 분류집단 0으로 분류한다. 또한 다음의 계산식 $0.8 \times x1 + 1.8 \times x2$의 값이 0.6보다 작으면 분류집단 1로 분류한다. 위의 두 가지 경우 모두 만족시키지 못하면 분류집단 2로 분류한다. 이 때 입력변수 $x3$와 $x4$는 분류에 영향을 주지 않았다.

	CHAID	CART	QUEST	BNN
1	0.0574	0.0738	0.0672	0.0131
2	0.0691	0.0852	0.0707	0.0113
3	0.0768	0.0735	0.0551	0.0050
4	0.0872	0.1141	0.0935	0.0111
5	0.0738	0.0789	0.0789	0.0274
6	0.0583	0.0686	0.0755	0.0086
7	0.0725	0.0843	0.1113	0.0152
8	0.0778	0.0927	0.0844	0.0099
9	0.0740	0.0740	0.0547	0.0113
10	0.0974	0.0927	0.0911	0.0128
11	0.0645	0.0697	0.0749	0.0070
12	0.0554	0.0675	0.0588	0.0052
13	0.0579	0.0727	0.0678	0.0182
14	0.0775	0.1066	0.0856	0.0081
15	0.0777	0.0744	0.0631	0.0097
16	0.0772	0.0856	0.0822	0.0067
17	0.0782	0.0749	0.0815	0.0083
18	0.0797	0.0813	0.0683	0.0065
19	0.0667	0.0889	0.0730	0.0032
20	0.1254	0.0603	0.0733	0.0098

오분류율 비교표

	평균	표준편차	최소값	최대값
BNN	.0104	.0054	.0032	.0274
CART	.0810	.0133	.0603	.1141
CHAID	.0752	.0158	.0554	.1254
QUEST	.0755	.0138	.0547	.1113

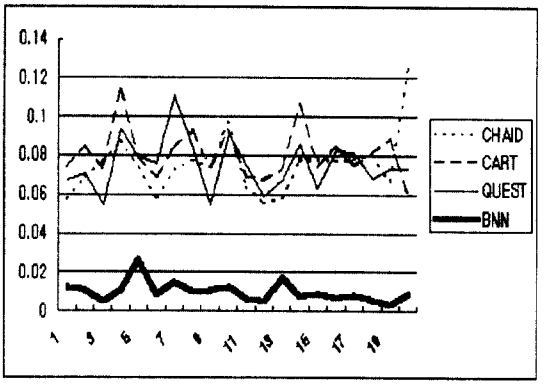

그림 3-1 : 분류결과도표

[그림 3-1]의 오분류율 분류결과도표에서와 같이 총20회 반복실험의 모든 경우에서 베이지안 신경망(BNN) 방법의 오분류율이 적었으며, 또한 전체 오분류율의 평균을 비교한

오분류율 비교표에서도 0.0104의 오분류율로 가장 분류를 잘했음을 알 수 있다.

3.2 IRIS DATA(총 자료수 150개)

자료설명 - 붓꽃의 특성치에 관한 자료로서 4개의 입력변수 $(x1, x2, x3, x4)$를 가진다. 입력변수가 되는 꽃받침의 길이 $(x1)$, 꽃받침의 넓이 $(x2)$, 꽃잎의 길이 $(x3)$, 꽃잎의 넓이 $(x4)$에 따라 붓꽃의 종류를 sectosa(분류집단 0), versicolor(분류집단 1),verginica(분류집단 2)의 세 집단으로 분류했다.

	CHAID	CART	QUEST	BNN
1	0.0882	0.0735	0.0735	0.0588
2	0.0741	0.0556	0.0370	0.0370
3	0.0690	0.0517	0.0517	0.0172
4	0.0656	0.0656	0.0820	0.0820
5	0.0847	0.0508	0.0508	0.0339
6	0.0294	0.0294	0.0294	0.0147
7	0.1833	0.0500	0.0500	0.0500
8	0.0833	0.0667	0.0333	0.0000
9	0.0149	0.0448	0.0000	0.0149
10	0.0938	0.0625	0.0781	0.0625
11	0.0862	0.0517	0.0517	0.0172
12	0.0462	0.0615	0.0154	0.0308
13	0.0702	0.0175	0.0351	0.0175
14	0.0984	0.0656	0.0656	0.0492
15	0.0952	0.0794	0.0317	0.0317
16	0.0615	0.0154	0.0308	0.0154
17	0.1667	0.0741	0.0741	0.0741
18	0.0508	0.0508	0.0169	0.0000
19	0.0946	0.0541	0.0541	0.0270
20	0.1364	0.0606	0.0303	0.0606

오분류율 비교표

	평균	표준편차	최소값	최대값
BNN	.0347	.0239	.0000	.0820
CART	.0541	.0172	.0154	.0794
CHAID	.0846	.0408	.0149	.1833
QUEST	.0446	.0225	.0000	.0820

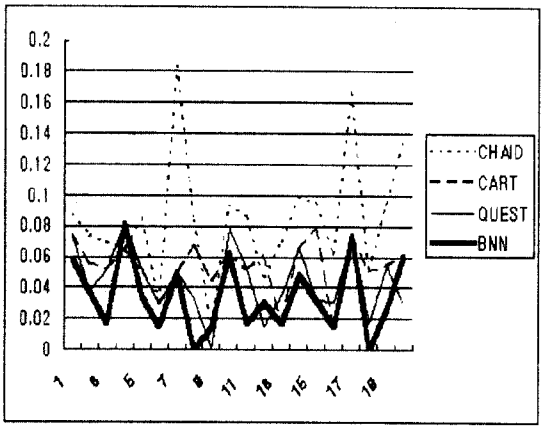

그림 3-2 : 분류결과도표

[그림 3-2]의 분류결과도표에서와 같이 총20회 반복실험의 모든 경우에서 베이지안 신경망(BNN) 방법의 오분류율이 작게 나타나지는 않았지만, 전체 오분류율의 평균을 비교한 오분류율 비교표에서 0.0347의 오분류율로 비교적 분류를 잘했음을 알 수 있다.

3.3 CREDIT DATA(총 자료수 323개)

자료설명 - 고객의 신용정도에 관한 자료로서 4개의 입력변수 $(x1, x2, x3, x4)$를 가진다. 입력변수가 되는 직위 $(x1 - 1.2.3.4.5.)$, 급여형태 $(x2 - 1.2.)$, 나이 $(x3 - 1.[< 25]2.[25 - 35]3.[35 <])$, 아멕스카드소지여부 $(x4 - 1.2.)$에 따라 개인의 신용여부를 불량등급(분류집단 0), 우량등급(분류집단 1) 의 두 집단으로 분류했다.

	CHAID	CART	QUEST	BNN
1	0.1119	0.1119	0.1194	0.0970
2	0.1261	0.1261	0.1261	0.0924
3	0.1552	0.1466	0.1552	0.1379
4	0.1862	0.1862	0.1862	0.1655
5	0.1858	0.1327	0.2035	0.1150
6	0.1608	0.2168	0.2168	0.1538
7	0.1194	0.1194	0.1269	0.1269
8	0.1269	0.1269	0.1269	0.1940
9	0.1395	0.1163	0.1395	0.1473
10	0.1181	0.1181	0.1181	0.1102
11	0.1508	0.1508	0.1746	0.1349
12	0.1504	0.1353	0.1504	0.1504
13	0.2030	0.2030	0.2030	0.1504
14	0.1221	0.1221	0.1374	0.1145
15	0.1888	0.1888	0.2098	0.2168
16	0.1367	0.1367	0.1439	0.1295
17	0.1391	0.1391	0.1652	0.1478
18	0.1628	0.1628	0.1860	0.1628
19	0.1214	0.1143	0.1857	0.1214
20	0.1071	0.1071	0.1286	0.1214

오분류율 비교표

	평균	표준편차	최소값	최대값
BNN	.1395	.0306	.0924	.2168
CART	.1431	.0322	.1071	.2168
CHAID	.1456	.0283	.1071	.2030
QUEST	.1602	.0330	.1181	.2188

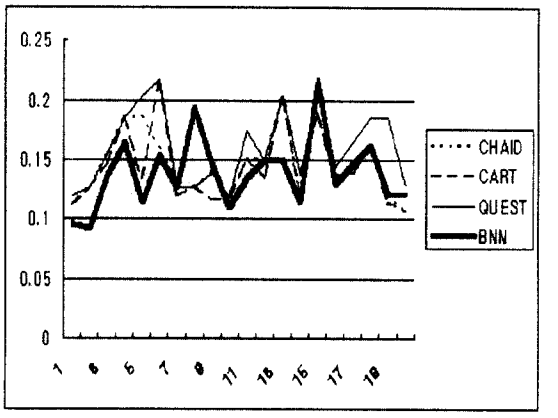

그림 3-3 : 분류결과도표

[그림 3-3]의 분류결과도표에서와 같이 총20회 실험의 모든 경우에서 베이지안 신경망(BNN)방법의 오분류율이 작게 나타나지는 않았지만, 전체 오분류율의 평균을 비교한 오분류율 비교표에서 0.1395의 오분류율로 근소한 차이지만 비교적 분류를 잘했음을 알 수 있다.

3.4 DIGIT DATA(총 자료수 200개)

자료설명 - 고장난 전자계산기로부터 얻어진 자료로서 7개의 입력변수 $(x1, x2, x3, x4, x5, x6, x7)$를 가지며 숫자 $0-9$의 10개의 집단으로 분류했다.

	CHAID	CART	QUEST	BNN
1	0.4598	0.2989	0.2989	0.3103
2	0.4110	0.2192	0.4110	0.2466
3	0.4416	0.3377	0.4286	0.2727
4	0.4819	0.3253	0.3253	0.2771
5	0.4930	0.3099	0.3239	0.2676
6	0.4149	0.3617	0.2553	0.2660
7	0.4337	0.3253	0.3253	0.2892
8	0.5529	0.4000	0.5529	0.3647
9	0.5698	0.3837	0.4419	0.3023
10	0.5412	0.4235	0.4118	0.4118
11	0.4875	0.3500	0.3625	0.3500
12	0.4483	0.3563	0.2989	0.2759
13	0.4545	0.2987	0.4545	0.2727
14	0.4872	0.2692	0.4872	0.3077
15	0.5287	0.3448	0.5172	0.3103
16	0.3523	0.3295	0.3409	0.2614
17	0.5132	0.2895	0.3947	0.2632
18	0.5263	0.3289	0.4737	0.3026
19	0.4330	0.3299	0.5052	0.3196
20	0.5059	0.2941	0.5176	0.3765

오분류율 비교표

	평균	표준편차	최소값	최대값
BNN	.3024	.0437	.2466	.4118
CART	.3288	.0460	.2192	.4235
CHAID	.4768	.0546	.3523	.5698
QUEST	.4064	.0868	.2553	.5529

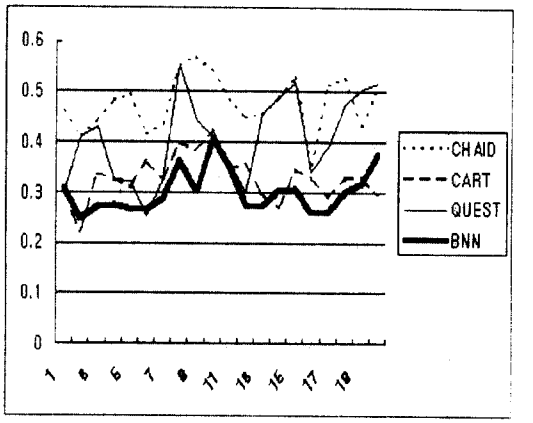

그림 3-4 : 분류결과도표

[그림 3-4]의 분류결과도표에서와 같이 총20회 실험의 대부분 경우에서 베이지안 신경망(BNN) 방법의 오분류율이 작게 나타나는 것을 알 수 있으며, 전체 오분류율의 평균을 비교한 오분류율 비교표에서도 0.3024의 오분류율로 비교적 분류를 잘했음을 알 수 있다.

3.5 WAVE DATA(총 자료수 5000개)

자료설명 - Wave형태에 관한 자료(Breiman et. Al.,1984)로서 21개의 연속적인 입력변수 ($x1 - x21$)를 가진다. 입력변수에 따라 분류집단 0, 분류집단 1, 분류집단 2의 세 집단으로 분류했다.

	CHAID	CART	QUEST	BNN
1	0.3561	0.2500	0.2629	0.1414
2	0.3633	0.2480	0.2600	0.1342
3	0.3511	0.2517	0.2596	0.1489
4	0.3528	0.2594	0.2569	0.1439
5	0.3752	0.2359	0.2557	0.1286
6	0.3739	0.2382	0.2608	0.1291
7	0.4397	0.2493	0.2765	0.1266
8	0.3371	0.2438	0.2554	0.1282
9	0.3877	0.2593	0.2511	0.1228
10	0.3449	0.2318	0.2586	0.1247
11	0.3874	0.2472	0.2472	0.1359
12	0.3924	0.2300	0.2350	0.1326
13	0.3554	0.2544	0.2444	0.1200
14	0.3563	0.2271	0.2627	0.1227
15	0.3908	0.2384	0.2649	0.1349
16	0.3307	0.2453	0.2757	0.1439
17	0.3499	0.2409	0.2388	0.1168
18	0.3483	0.2461	0.2587	0.1309
19	0.3777	0.2308	0.2520	0.1248
20	0.4019	0.2469	0.2715	0.1284

오분류율 비교표

	평균	표준편차	최소값	최대값
BNN	.1310	.0086	.1168	.1489
CART	.2437	.0094	.2271	.2594
CHAID	.3686	.0261	.3307	.4397
QUEST	.2574	.0108	.2350	.2765

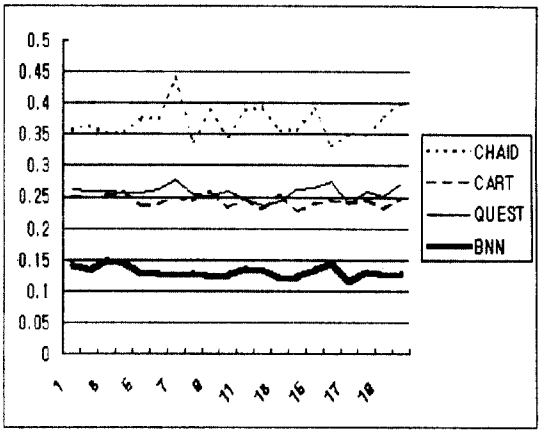

그림 3-5 : 분류결과도표

[그림 3-5]의 분류결과도표에서와 같이 총20회 실험의 모든 경우에서 베이지안 신경망(BNN) 방법의 오분류율이 적었으며, 또한 전체 오분류율의 평균을 비교한 오분류율 비교표에서도 0.1310의 오분류율로 가장 분류를 잘했음을 알 수 있다.

제 4 절 결론

모든 자료에서 베이지안 신경망 방법이 분류알고리즘인 CHAID, CART, QUEST의 방법들보다 정확도면에서 좋음을 보였다. 특히 연속형 자료인 3WAY DATA나 WAVE DATA에서처럼 자료의 양이 많은 경우와 입력변수의 자료가 베이지안에서의 사후확률분포를 변경시킬 만큼 구체적인 자료를 가질 때 다른 분류알고리즘인 CHAID, CART, QUEST의 방법들과의 오분류율 비교에서 더 큰 차이를 보였다. 또한 CREDIT DATA나 DIGIT DATA에서와 같이 입력자료의 형태가 범주형 자료일때의 분석결과도 베이지안 신경망 방법이 비교적 좋게 나왔다. 이와 같이 베이지안 신경망 방법의 사용이 정확도 면에서 좋으나 다른 방법에 비해 분류에 많은 시간이 소요되어 이 부분에 대한 개선이 필요하다고 생각된다.

참 고 문 헌

1. Bishop , C. M. (1995), Neural Networks for Pattern Recognition, Oxford University Press.

2. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), Classification and Regression Trees, Belmont, California:Wadsworth.

3. Bridle, J, S. (1989), Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, in F. Fouglemann-Soulie and J.Heault (editors) Neuro-computing: Algorithms, Architectures and Applications, New York: Springer-Verlag.

4. Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), Hybrid Monte Carlo, Physics Letters B, vol. 195, pp. 216-222.

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

6. Mackay, D. J. C. (1994a), Bayesian non-linear modeling for the energy prediction competition, ASHRAE Transactions, vol. 100, pt. 2, pp. 1053-1062.

7. Mackay, D. J. C. (1994b), Hyperparameters: Optimise, or integrate out?, in G. Heidbreder, editor, Maximum Entropy and Bayesian Methods, Santa Barbara, 1993, Dordrecht: Kluwer.

8. Neal, R. M. (1996), Bayesian Learning for Neural Networks, Springer.

9. Quinlan, R. (1993), Combining instance-based and model-based learning, Machine Learning: Proceedings of the Tenth International Conference, Amherst, Massachusetts, 1993, Morgan Kaufmann.

10. Ripley, B. D. (1996), Pattern Recognition and Neural Networks, Cambridge University Press.

11. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986a), Learning representations by back-propagating errors, Nature, vol. 323, pp. 533-536.

12. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b), Learning internal representations by error propagation, in D. E. Rumelhart and J. L. McClelland (editors) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations, Cambridge, Massachusetts: MIT Press.

13. 최종후, 한상태, 강현철, 김은식 (1998), AnswerTree를 이용한 데이터 마이닝 의사결정나무분석, SPSS 아카데미.

A Classification Analysis using Bayesian Neural Network

Jinsoo Hwang[5] · SeongYong Choi[6] · HongSuk Jun[7]

Abstract

There are several algorithms for classification in modeling relations, patterns, and rules which exist in data. We learn to classify objects on the basis of instances presented to us, not by being given a set of classification rules. The Bayesian learning uses the probability distribution to express our knowledge about unknown parameters and update our knowledge by the law of probability as the evidence gathered from data. Also, the neural network models are designed for predicting an unknown category or quantity on the basis of known attributes by training. In this paper, we compare the misclassification error rates of Bayesian Neural Network method with those of other classification algorithms, CHAID, CART, and QUEST using several data sets.

Key Words and Phrases: ARD, Classification, Bayesian Neural Network, Hybrid Monte Carlo

[5]Associate Professor, Department of Statistics, Inha University, 402-751, Inchon, Korea
[6]Candidate for the Ph. D, Department of CSE, Inha University, 402-751, Inchon, Korea
[7]Professor, Department of Statistics, Inha University, 402-751, Inchon, Korea