응용통계연구 (2008)
21(5), 825-833

정규 혼합분포를 이용한 준지도 학습

최병정[1] · 채윤석[2] · 최우영[3] · 박창이[4] · 구자용[5]

[1]고려대학교 통계학과; [2]SAS Korea, 컨설턴트; [3]SAS Korea, 컨설턴트;
[4]서울시립대학교 통계학과; [5]고려대학교 통계학과

(2008년 3월 접수, 2008년 7월 채택)

요 약

혼합모형을 이용한 판별분석은 다중 분류문제를 해결하는데 유용한 방법으로서 준지도 학습으로 확장될 수 있다. 본 논문에서는 정규 혼합분포를 이용한 준지도 학습 방법에서 혼합 모형의 하위 구성요소 개수 선택 기준을 연구하고자 한다. 하위 구성요소 선택 기준으로서 베이지안 정보량을 사용하였고 모의실험을 통해 이 방법의 유용성을 규명하였다.

주요용어: 베이지안 정보량, 분류, 밀도 추정, EM 알고리즘, 정규 혼합분포.

1. 서론

데이터마이닝에서 주요 연구 분야의 하나인 학습(Learning)은 크게 입력 및 출력변수의 값이 쌍으로 관측되어 출력변수가 학습을 지도하는 역할을 하는 지도학습(Supervised Learning)과 출력변수가 없는 경우인 비지도학습(Unsupervised Learning)으로 나뉘어진다. 입력 및 출력변수가 모두 관측되는 자료를 라벨이 있는 자료(Labeled Data)라 하고 출력변수가 없는 자료를 라벨이 없는 자료(Unlabeled Data)라 한다. 전통적인 분류 기법들은 라벨이 있는 자료만을 다루는 지도학습으로 볼 수 있다. 그러나 출력변수의 값을 얻기 위해서는 전문인력이 하나씩 자료를 검토해야 하므로 많은 시간과 비용이 소요되지만 입력변수의 값은 관측하기는 상대적으로 쉬운 경우를 흔히 볼 수 있다. 이러한 경우는 예컨대 텍스트 마이닝에서 다량의 웹 문서를 분류한다든지 생물학에서 유전자의 기능을 분류할 때 흔히 발생한다.

최근 많이 연구되는 준지도학습(Semi-Supervised Learning)은 분류문제에서 라벨이 있는 자료 뿐 아니라 라벨이 없는 자료를 이용하여 분류의 정확도를 향상시키는 것을 그 목적으로 한다. 앞서 설명한 배경하에서 준지도학습은 라벨을 얻기 위한 비용을 줄이는 동시에 더 정확한 분류 결과를 줄 수도 있기 때문에 이에 대한 연구는 이론적인 측면 뿐만아니라 현실적인 측면에서도 의의가 크다고 할 수 있다. 준지도 학습에 대한 보다 자세한 소개 및 여러 가지 방법론에 대한 설명은 Zhu (2005)에 나와 있다.

이 연구는 고려대학교 특별 연구비에 의하여 수행되었음.
[1]교신저자: (135-839) 서울시 강남구 대치4동 889-11 대치B/D 9F, SAS Korea, 고려대학교 통계학과, 박사.
E-mail: hessian@korea.ac.kr, byoung-jeong.choi@sas.com
[2](135-839) 서울시 강남구 대치4동 889-11 대치B/D 9F, SAS Korea, 컨설턴트.
E-mail: youn-seok.chae@sas.com
[3](135-839) 서울시 강남구 대치4동 889-11 대치B/D 9F, SAS Korea, 컨설턴트.
E-mail: woo-young.choi@sas.com
[4](130-743) 서울 동대문구 전농동 90, 서울시립대학교 자연대학 통계학과, 조교수. E-mail: park463@uos.ac.kr
[5](136-701) 서울 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수. E-mail: jykoo@korea.ac.kr

Hastie와 Tibshirani (1996)는 정규 혼합모형에 기반한 혼합판별분석(Mixture Discriminant Analysis: MDA)을 제안하였다. 혼합 판별분석은 혼합 밀도의 구성 요소의 수를 적절히 선택하면 다양한 형태의 확률 밀도 함수를 근사할 수 있으며, Dempster 등 (1977)이 제안한 기대화-최대화(Expectation-Maximization: EM) 알고리즘으로 모수의 추정이 쉽게 구현될 수 있다는 장점이 있다. Nigam 등 (2000)은 연속형 변수를 이산화하여 혼합다항분포를 학습하는 소위 단순 베이즈 분류방법(Naive Bayes Classifier)을 이용한 준지도 학습을 제안하였다. Halbe와 Aladjem (2005)는 지도 학습에서 여러가지 형태의 공분산행렬에 대하여 정규 혼합모형의 하위 구성요소 개수 선택 기준으로 베이지안 정보량(Bayesian Information Criterion: BIC)를 제안하였다.

본 논문에서는 Halbe와 Aladjem (2005)의 결과를 준지도 학습문제에 적용하고자 한다. 특히 관측된 자료의 개수에 비해 입력변수의 개수가 많은 경우에 발생할 수 있는 과모수화(Over-parameterization) 문제를 완화시키기 위하여 Halbe와 Aladjem (2005)에서도 연구된 바 있는 대각행렬 형태의 공분산 행렬을 갖는 정규 혼합모형만을 고려하였다. 편의상 이를 준지도 혼합 판별분석(Semi-Supervised Mixture Discriminant Analysis: SS-MDA)이라 부르기로 한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 SS-MDA를 소개하고 3절에서는 Waveform 자료와 혼합 분포에서의 모의실험을 통해 SS-MDA에서의 하위 구성요소 개수를 선택하는 기준인 BIC가 효율적임을 보일 것이다. 마지막으로 4절에서는 본 논문의 결과를 요약하며 추후 연구 방향을 제시한다.

2. 준지도 혼합 판별 분석

입력변수와 출력변수가 각각 $\boldsymbol{X} = (X^1, \ldots, X^p) \in \mathbb{R}^p$와 $Y \in \mathcal{K} = \{1, \ldots, K\}$로 나타내어지는 다중 분류 문제를 생각해 보자. 훈련표본 $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, N$은 미지의 입력변수과 출력변수의 결합분포로부터 서로 독립적으로 생성된다고 하자. 여기서 $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^p)'$이다. 본 논문에서는 준지도 학습을 고려하므로 $i = M+1, \ldots, N$ $(M < N)$에 대해서는 출력변수의 값이 관측되지 않는다고 가정한다.

2.1. SS-MDA에 대한 소개

SS-MDA에서는 MDA와 동일하게 조건부 확률 밀도 함수 $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \,|\, Y = k)$, $k = 1, \ldots, K$가 동일한 공분산 행렬 Σ를 갖는 정규 혼합모형이라 가정한다. 즉, 각 클래스 k를 가상적인 R_k개의 하위 구성요소 c_{kr}, $r = 1, \ldots, R_k$로 나누고, 각 하위 구성요소 c_{kr}은 평균 벡터가 $\boldsymbol{\mu}_{kr}$이고 공분산 행렬이 Σ인 다변량 정규분포를 따른다고 가정한다. 이때 변수의 개수가 많고 자료의 개수가 충분하지 않은 경우에 생길 수 있는 과모수화 문제를 피하기 위하여 대각행렬 형태의 공분산 행렬만을, 즉 $\Sigma = \mathrm{diag}(\tau_1, \ldots, \tau_p)$을 고려하기로 한다.

클래스 k에 대한 사전 확률을 Π_k로 클래스 k 내에서 r번째 하위 구성요소 혼합 확률을 π_{kr}로 나타내면 클래스 k의 혼합 밀도 함수는 다음과 같다.

$$m_k(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{r=1}^{R_k} \pi_{kr} |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{D_\Sigma^2(\boldsymbol{x}, \boldsymbol{\mu}_{kr})}{2} \right\},$$

여기서 $\sum_{k=1}^{K} \Pi_k = 1$, $\sum_{r=1}^{R_k} \pi_{kr} = 1$, $D_\Sigma(\boldsymbol{x}, \boldsymbol{\mu}) = \left[(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right]^{1/2}$은 $\boldsymbol{x}$와 $\boldsymbol{\mu}$간의 마할라노비스 거리 그리고 $\boldsymbol{\theta} = (\Pi_1, \ldots, \Pi_k,\ \pi_{11}, \ldots, \pi_{kr},\ \boldsymbol{\mu}_{11}^T, \ldots, \boldsymbol{\mu}_{kr}^T,\ \tau_1, \ldots, \tau_p)$를 나타낸다.

그러면 SS-MDA는 실제로 관측되는 자료 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_M, y_M)$, $\boldsymbol{x}_{M+1}, \ldots, \boldsymbol{x}_N$를 이용하여 모수 $\boldsymbol{\theta}$를 추정하는 문제가 된다. 이에 대한 가능도 함수는 다음과 같이 표현할 수 있다.

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{M} \Pi_{y_i} m_{y_i}(\boldsymbol{x}_i;\, \boldsymbol{\theta}) \prod_{i=M+1}^{N} \sum_{k=1}^{K} \Pi_k m_k(\boldsymbol{x}_i;\, \boldsymbol{\theta}).$$

수치적으로 $L(\boldsymbol{\theta})$를 최대화하는 추정치 $\hat{\boldsymbol{\theta}}$를 구하는 문제는 복잡하므로 EM 알고리즘 이용하기로 한다. EM 알고리즘에 의해 유도된 SS-MDA의 알고리즘은 2.2절에서 자세히 설명한다. 그 유도 과정은 그리 어렵지 않으므로 생략하기로 한다.

일단 SS-MDA 알고리즘에 의해 추정치 $\hat{\boldsymbol{\theta}}$을 구하면 새로운 관측값 $\boldsymbol{x}$에 대하여 다음과 같이 클래스 조건부 확률의 추정값을 베이즈 공식을 이용하여 얻을 수 있다.

$$\hat{\mathbb{P}}(Y = k \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\sum_{r=1}^{R_k} \hat{\Pi}_k \hat{\pi}_{kr} \exp\left\{-\frac{D_{\hat{\Sigma}}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{kr})}{2}\right\}}{\sum_{k'=1}^{K} \sum_{r'=1}^{R_{k'}} \hat{\Pi}_{k'} \hat{\pi}_{k'r'} \exp\left\{-\frac{D_{\hat{\Sigma}}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{k'r'})}{2}\right\}}.$$

이때 새로운 관측값 $\boldsymbol{x}$의 클래스 예측값은 $\arg\max_{k=1,\ldots,K} \hat{\mathbb{P}}(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$으로 주어진다.

2.2. SS-MDA 알고리즘

다음은 EM 알고리즘에 의해 유도된 SS-MDA의 알고리즘이다.

1. 초기치 설정

   클래스 k에 대한 사전 확률인 Π_k는 클래스가 관측된 완전 정보만을 이용하여 클래스별 비율로 초기치를 설정한다. 그리고 클래스가 k이고 하위 구성요소가 r인 경우에 대한 혼합 확률 π_{kr}은 r개의 하위 구성요소에 대해 클래스 k 내에서 동일한 비율인 $1/R_k$로 초기치를 설정한다. 즉,

   $$\Pi_k^{(0)} = \frac{\sum_{i=1}^{M} I(y_i = k)}{M}, \qquad \pi_{kr}^{(0)} = \frac{1}{R_k}.$$

   그리고 평균 벡터인 $\boldsymbol{\mu}_{kr}^{(0)}$들은 K-평균군집에서 구한 중심점(centroid)들을 이용하고, 공분산 $\Sigma^{(0)}$는 다음과 같이 초기치를 설정한다.

   $$\Sigma^{(0)} = \frac{1}{N} \sum_i \sum_k \sum_r \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr}^{(0)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr}^{(0)}\right)^T.$$

2. 기대화 과정

   기대화 과정은 다음과 같이 정의된 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$를 계산하는 과정이다.

   $$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(\log \Pi_k + \log m_k\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right) \gamma_k\left(\mathcal{O}_i, \boldsymbol{\theta}^{(t)}\right),$$

   여기서,

   $$\mathcal{O}_i = \begin{cases} (\boldsymbol{x}_i, y_i), & 1 \le i \le M, \\ \boldsymbol{x}_i, & M+1 \le i \le N \end{cases}$$

은 관측된 정보를 나타내고

$$\gamma_k(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}) = \begin{cases} 1, & i \le M,\, y_i = k, \\ 0, & i \le M,\, y_i \ne k, \\ \dfrac{\Pi_k^{(t)} m_k(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{K} \Pi_k^{(t)} m_k(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})}, & i \ge M+1. \end{cases}$$

3. 최대화 과정

최대화 과정은 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$를 최대화하는 $\boldsymbol{\theta}^{(t+1)}$을 구하는 과정으로 $\boldsymbol{\theta}^{(t+1)}$은 다음과 같다.

$$\begin{aligned}
\Pi_k^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right), \\
D_{\Sigma}^{(t+1)}\left(\boldsymbol{x}_i,\, \boldsymbol{\mu}_{kr}\right) &= \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr}^{(t)}\right)^T \left(\Sigma^{(t)}\right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr}^{(t)}\right), \\
p^{(t+1)}\left(c_{kr} \mid \boldsymbol{x}_i, k\right) &= \frac{\pi_{kr}^{(t)} \exp\left\{-\dfrac{D_{\Sigma}^{(t+1)}\left(\boldsymbol{x}_i,\, \boldsymbol{\mu}_{kr}\right)}{2}\right\}}{\sum_{r'} \pi_{kr'}^{(t)} \exp\left\{-\dfrac{D_{\Sigma}^{(t+1)}\left(\boldsymbol{x}_i,\, \boldsymbol{\mu}_{kr'}\right)}{2}\right\}}, \\
\pi_{kr}^{(t+1)} &= \frac{\sum_i p^{(t+1)}\left(c_{kr} \mid \boldsymbol{x}_i,\, k\right) \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right)}{\sum_i \sum_{r'} p^{(t+1)}\left(c_{kr'} \mid \boldsymbol{x}_i,\, k\right) \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right)}, \\
\boldsymbol{\mu}_{kr}^{(t+1)} &= \frac{\sum_i p^{(t+1)}\left(c_{kr} \mid \boldsymbol{x}_i,\, k\right) \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right) \boldsymbol{x}_i}{\sum_i p^{(t+1)}\left(c_{kr} \mid \boldsymbol{x}_i, k\right) \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right)}, \\
{\tau_j}^{(t+1)} &= \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{r=1}^{R_k} p^{(t+1)}\left(c_{kr} \mid \boldsymbol{x}_i,\, k\right) \left(\boldsymbol{x}^j - \boldsymbol{\mu}_{kr}^{j(t+1)}\right)^2 \gamma_k \left(\mathcal{O}_i,\, \boldsymbol{\theta}^{(t)}\right).
\end{aligned}$$

4. $L(\boldsymbol{\theta}^{(t)})$가 수렴할 때까지 기대화 과정과 최대화 과정을 반복하며 수렴조건이 성립할 때의 $\boldsymbol{\theta}^{(t)}$를 $\hat{\boldsymbol{\theta}}$로 정의한다.

2.3. 최적 하위 구성요소 개수의 선택 기준

혼합 모형에서 하위 구성요소 개수 R_k가 너무 작으면 적합도가 결여되어 추정의 정확도가 떨어지게 되고 R_k가 지나치게 크면 훈련 자료에 과적합되는 문제가 발생할 수 있다. 따라서 최적의 하위 구성요소 개수의 선택 방법이 필요한데, 본 연구에서는 그 기준으로서

$$\mathrm{BIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + \ln(N) \left\{ \sum_{k=1}^{K} R_k(p+1) + K + p \right\}$$

를 이용하였다. 여기서 추정하는 모수는 $\boldsymbol{\theta} = (\Pi_1, \ldots, \Pi_k,\, \pi_{11}, \ldots, \pi_{kr},\, \boldsymbol{\mu}_{11}^T, \ldots, \boldsymbol{\mu}_{kr}^T,\, \tau_1, \ldots, \tau_p)$이므로 모형의 복잡도(model complexity)항이 $\sum_{k=1}^{K} R_k(p+1) + K + p$가 된다. 최소 BIC를 갖는 $\{\hat{R}_k\}$을 최적의 하위 구성요소 개수로 선택하고자 한다.

표 3.1. WAVEFORM 자료에서 MDA와 SS-MDA의 오분류율 비교

라벨이 있는 자료의 수	MDA	SS-MDA-L	SS-MDA
50	0.3110 (0.0035)	0.2186 (0.0032)	0.2060 (0.0039)
100	0.2339 (0.0024)	0.1823 (0.0017)	0.1879 (0.0024)
200	0.2003 (0.0017)	0.1683 (0.0015)	0.1669 (0.0016)
300	0.1816 (0.0014)	0.1598 (0.0014)	0.1626 (0.0014)
500	0.1625 (0.0012)	0.1540 (0.0013)	0.1535 (0.0011)

1. 하위 구성요소 개수의 최대값 $R_{\max}$를 정한다.
2. $k = 1, \ldots, K$에 대하여 $R_k = 1$로 설정된 초기 모형에 주어진 자료를 적합한 후 BIC를 계산한다.
3. 모든 $k = 1, \ldots, K$에 대하여 각 클래스의 개수가 $(R_1, \ldots, R_{k-1}, R_k + 1, R_{k+1}, \ldots, R_K)$인 모형을 설정하고 BIC를 계산한다.
4. 모든 $k = 1, \ldots, K$에 대하여 $R_k = R_{\max}$이거나 계산된 BIC가 이전 단계에서 계산된 값보다 증가하는 경우 탐색을 멈추고 그 때의 하위 구성요소 조합을 $\{\hat{R}_k\}$으로 선택한다.

3. 모의실험

3.1. Waveform data

첫 번째 모의실험은 Breiman 등 (1984)과 Hastie와 Tibshirani (1996) 등에서 사용된 Waveform 자료를 이용하여 수행하였다. Waveform 모의실험에서는 3개의 클래스와 21개의 입력변수가 생성된다.

$$X^j = \begin{cases} Uh_1(j) + (1-U)h_2(j) + \epsilon_j, & \text{클래스 1,} \\ Uh_1(j) + (1-U)h_3(j) + \epsilon_j, & \text{클래스 2,} \\ Uh_2(j) + (1-U)h_3(j) + \epsilon_j, & \text{클래스 3,} \end{cases}$$

여기서 $j = 1, 2, \ldots, 21$이고, $U \sim U(0, 1)$, $\epsilon_j \sim N(0, 1)$이며, h_l은 다음과 같이 정의된 시프트된 삼각 웨이브폼(shifted triangular waveforms)이다.

$$\begin{aligned} h_1(j) &= \max(6 - |j - 11|, 0), \\ h_2(j) &= h_1(j - 4), \\ h_3(j) &= h_1(j + 4). \end{aligned}$$

보다 자세한 설명은 Breiman 등 (1984)를 참조하기 바란다. 이 실험의 목적은 하위 구성요소의 개수를 선택할 수 있는 SS-MDA와 그러한 기능이 없는 MDA와의 비교를 통해 하위 구성요소 개수 선택이 예측력 향상에 중요할 수 있음을 보이는 것이다. 공정한 비교를 위해 훈련 자료에서 출력 변수의 값에 결측이 생기지 않도록 하였다. 즉, 라벨이 있는 자료만을 훈련자료로 사용하였다. 시험자료의 크기는 1000으로 고정시키고 훈련자료는 라벨이 없는 자료와 라벨이 있는 자료를 합하여 사용하였다. 여기서 라벨이 없는 자료는 500으로 고정 시켰으며 라벨이 있는 자료는 50, 100, 200, 300, 500으로 하여 MDA, SS-MDA-L 그리고 SS-MDA의 오분류율을 비교하였다. SS-MDA-L은 훈련에 라벨이 있는 자료만을 사용한 것으로 BIC에 의한 하위 구성요소 개수선택의 효과와 라벨이 없는 자료의 효과를 비교하기 위한 것이다. 오분류율의 변동성을 파악하기 위하여 이 과정을 100번 반복하였다.

표 3.1은 MDA, SS-MDA-L, SS-MDA를 적합시킨 후 시험 자료에서 구한 오분류율의 평균과 표준오차를 보여준다. MDA, SS-MDA-L 그리고 SS-MDA 모두 라벨이 있는 자료(훈련자료)의 크기가 커질수

표 **3.2.** 혼합 분포 자료에서 최소 BIC값을 갖는 하위 구성요소 개수의 선택 빈도

R_1	R_2	MCAR-L	MCAR	MAR-L	MAR
2	1			13	
2	2	100	100	26	93
2	3			5	
3	1				6
3	2			11	1
3	3			30	
3	4			3	
4	3			5	
4	4			7	

록 오분류율이 낮아짐을 알 수 있다. 이는 자료의 크기가 클수록 정보의 양이 많아지기 때문에 모형의 예측 정확도가 높아지는 것이다. 또한 자료의 크기가 커짐에 따라 오분류율의 감소폭이 둔화됨을 볼 수 있다. 따라서 자료 개수가 일정 수준을 넘어가면 더는 예측력 향상에 도움이 되지 않음을 알 수 있다. 또한 하위 구성요소 개수를 선택하는 SS-MDA-L이 하위 구성요소 개수 선택 기능이 없는 MDA에 비해 예측력이 향상됨을 볼 수 있다. 특히 자료의 수가 작을 때는 그 차이가 두드러지고 자료의 크기가 커질수록 그 차이가 줄어드는 것을 확인할 수 있다. 따라서 자료의 개수가 많지 않은 경우에는 SS-MDA에서의 하위 구성요소 개수 선택에 의한 모형 최적화가 예측력 향상에 도움을 줌을 알 수 있다. SS-MDA-L과 SS-MDA를 비교하면 오분류율이 비슷하므로 라벨이 없는 자료의 효과는 거의 없다고 볼 수 있다. 그 이유는 라벨이 없는 자료가 MCAR(missing completely at random) 가정하에서 생성되기 때문이다. 즉 완전히 랜덤하게 결측이 되는 상황에서는 라벨이 없는 자료를 사용하는 준지도 학습 방법은 지도 학습 방법에 비해 예측력의 향상을 기대하기 어렵다.

3.2. 혼합 분포 자료

BIC가 하위 구성요소 개수를 선택하는데 있어서 좋은 기준임을 보이기 위하여 다음과 같은 모의실험을 하였다. 이진 분류 문제를 고려하였고 클래스 1의 중심점이 각각 (1, 1)과 (3, 3)이고 클래스 2의 중심점은 (1, 2.6)과 (3, 1)이며 diag(0.7, 0.7)를 공분산 행렬로 갖는 정규 혼합모형을 고려하였다. 이 모형으로부터 2000개의 라벨이 있는 자료를 생성하여 1000개는 훈련자료로 나머지 1000개는 시험자료로 사용하였다.

준지도 학습 자료를 생성하기 위해서 우선 1000개의 라벨이 있는 원래의 훈련자료 중 MCAR 가정하에서는 랜덤하게 선택된 500개의 자료의 출력 변수를 제거하였고 MAR(missing at random) 가정하에서는 $x_2 - x_1 \geq 0.4$ 조건을 만족시키는 자료의 출력 변수의 값을 제거하였다. 이렇게 생성된 준지도 학습 자료에 대하여 SS-MDA를 적용하였고 자료 생성 및 SS-MDA 적합의 전 과정을 100번 반복하였다. 각 반복마다 BIC를 최소로 하는 하위 구성요소 개수의 조합을 찾아 그 선택 빈도를 구하였고 결과는 다음과 같다. 여기서 R_1과 R_2는 각 클래스의 하위 구성요소의 개수를 나타낸다.

표 3.2에서 MCAR, MCAR-L, MAR, MAR-L은 각각 MCAR가정하에서 생성된 1000개의 준지도 학습자료 전체, MAR가정하에서 라벨이 있는 500개의 자료, MAR가정하에서 생성된 1000개의 준지도 학습자료 전체, MAR가정하에서 라벨이 있는 $x_2 - x_1 < 0.4$를 만족시키는 자료를 SS-MDA에 적용한 결과이다. 100회 반복 중 MAR-L을 제외한 나머지에서는 최소 BIC를 가지는 하위 구성요소 개수의 조합이 $R_1 = 2$, $R_2 = 2$에서 대부분 관측되었다. MCAR가정하에서는 Waveform 자료에서 관측된 것처

표 3.3. 혼합 분포 자료에서 BIC에 의해 선택된 모형에 의한 시험 자료 오분류율

MCAR-L	MCAR	MAR-L	MAR
0.2250 (0.0015)	0.2249 (0.0014)	0.3336 (0.0027)	0.2359 (0.0035)

럼 라벨이 없는 자료를 이용하는 효과가 없는 것을 확인할 수 있다. MAR가정하에서는 라벨이 있는 자료만을 학습에 사용한 MAR-L의 경우에 비해 라벨이 없는 자료를 학습에 사용한 MAR의 경우에 하위 구성요소 개수의 참값 $R_1 = 2$, $R_2 = 2$를 잘 찾아줌을 볼 수 있다. MCAR의 경우와는 달리 MAR의 경우에는 라벨이 있는 자료는 원자료에 비해 편의(bias)가 있으며 라벨이 없는 자료를 훈련에 사용하는 것이 자료의 편의를 줄여주기 때문으로 생각된다.

표 3.3은 각 반복 단계에서 결정된 최적 모형에서 시험자료 오분류율의 평균과 표준오차를 보여준다. 오분류율을 비교해 보면 하위 구성요소 개수의 참값을 잘 찾은 경우에 오분류율이 낮음을 알 수 있다. MCAR가정하에서는 MCAR-L과 MCAR의 오분류율이 비슷하며 MCAR가정하에서는 라벨이 없는 자료를 사용하는 MAR이 라벨이 있는 자료만 사용하는 MAR-L에 비해 오분류율이 낮음을 볼 수 있다. 이 실험 결과로부터 BIC가 최적 하위 구성요소 개수를 선택하는데 있어서 효과적인 기준임을 알 수 있고, MAR가정과 같이 자료에 편의가 있는 경우 라벨이 있는 자료와 라벨이 없는 자료를 동시에 이용하는 준지도 학습 방법이 라벨이 있는 자료만을 이용하는 지도학습 방법에 비해 예측력을 향상시켜줌을 알 수 있다.

4. 결론

본 논문은 다음과 같이 요약될 수 있다. 첫째, 혼합모형을 이용한 분류문제에서 하위 구성요소 개수의 적절한 선택은 예측력을 향상시킬 수 있음을 알 수 있다. 3절의 모의실험을 통해서 살펴본 것처럼 SS-MDA는 BIC를 이용하여 하위 구성요소의 개수를 선택함으로써 효과적으로 모형을 최적화 할 수 있다. 둘째, MAR의 경우처럼 출력변수의 값이 자료의 특정 영역에서 관측되지 않는 상황에서 발생할 수 있는 자료의 편의를 줄일 수 있다. 이러한 상황은 금융산업의 신용평점화(Credit Scoring)에서 기각 추론(Reject Inference)과 매우 흡사하며 이 분야에도 적용할 수 있을 것이라 생각된다.

추후 연구과제로는 다음과 같은 것을 고려할 수 있다. 첫째, SS-MDA에서는 대각행렬형태의 공분산 행렬을 사용하여 추정해야할 모수의 개수를 줄임으로써 자료의 개수가 적은 경우에 발생할 수 있는 과모수화 문제를 어느 정도 완화시켜 줄 수 있다. 그러나 경우에 따라서는 대각 공분산행렬의 사용은 비효율적일 수 있다. 따라서 가설검정에 의하여 혹은 자료에 따라 자동적으로 공분산 행렬의 형태를 선택하는 방법을 생각해 볼 수 있다. 둘째, SS-MDA에서는 정규 혼합분포를 기초로 하기 때문에 이산형 자료를 효율적으로 다룰 수 없는 단점이 있는데, 이 경우 단순 베이지안 가정을 통해 연속형과 이산형 변수를 모두 독립적으로 추정하는 방법을 고려할 수 있다. 금융의 신용평점화 뿐만 아니라 대부분의 평점화 분야에서 연속형 변수를 이산화 또는 그룹화하여 사용하는 사례가 많기 때문에 실용적인 측면에서도 그 연구에 의의가 있다.

참고문헌

Breiman, L., Fredman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees,* Wadsworth, Belmont.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Jornal of the Royal Statistical Society, Series B*, **39**, 1–38.

Halbe, Z. and Aladjem, M. (2005). Model-based mixture discriminant analysis—an experimental study, *Pattern Recognition*, **38**, 437–440.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian Mixtures, *Journal of the Royal Statististical Society, Serise B*, **58**, 158–176.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM, *Machine Learning*, **39**, 103–134.

Zhu X. (2005). *Semi-Supervised Learning Literature Survey*, Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Semi-Supervised Learning by Gaussian Mixtures

Byoung-Jeong Choi[1] · Youn-Seok Chae[2] · Woo-Young Choi[3] · Changyi Park[4] · Ja-Yong Koo[5]

[1]Dept. of Statistics, Korea University; [2]Consultant, SAS Korea; [3]Consultant, SAS Korea;
[4]Dept. of Statistics, University of Seoul; [5]Dept. of Statistics, Korea University

(Received March 2008; accepted July 2008)

Abstract

Discriminant analysis based on Gaussian mixture models, an useful tool for multi-class classifications, can be extended to semi-supervised learning. We consider a model selection problem for a Gaussian mixture model in semi-supervised learning. More specifically, we adopt Bayesian information criterion to determine the number of subclasses in the mixture model. Through simulations, we illustrate the usefulness of the criterion.

Keywords: BIC, classification, density estimation, EM algorithm, Gaussian mixture.

This research was supported by a Korea University Grant.

[1]Corresponding author: Doctor, Dept. of Statistics, Korea University. SAS Korea, 9F, Daechi B/D, 889-11, Daechi-dong, Gangnam-gu, Seoul 135-839, Korea.
E-mail: hessian@korea.ac.kr, byoung-jeong.choi@sas.com

[2]Consultant, SAS Korea, 9F, Daechi B/D, 889-11, Daechi-dong, Gangnam-gu, Seoul 135-839, Korea.
E-mail: youn-seok.chae@sas.com

[3]Consultant, SAS Korea, 9F, Daechi B/D, 889-11, Daechi-dong, Gangnam-gu, Seoul 135-839, Korea.
E-mail: woo-young.choi@sas.com

[4]Assistant Professor, Dept. of Statistics, University of Seoul, Jeonnong-dong 90, Seoul 130-743, Korea.
E-mail: park463@uos.ac.kr

[5]Professor, Dept. of Statistics, Korea University. Anam-dong 5-1, Seoul 136-701, Korea.
E-mail: jykoo@korea.ac.kr