주성분 분석과 나이브 베이지안 분류기를 이용한 퍼지 군집화 모형

전 성 해†

요 약

자료의 표현에서 군집화는 주어진 데이터를 서로 유사한 개체들끼리 몇 개의 집단으로 묶는 작업을 수행한다. 군집화의 유사도 결정 측도는 많은 연구들에서 매우 다양한 것들이 사용되었다. 하지만 군집화 결과의 성능 측정에 대한 객관적인 기준 설정이 어렵기 때문에 군집화 결과에 대한 해석은 매우 주관적이고, 애매한 경우가 많다. 퍼지 군집화는 이러한 주관적인 군집화 문제에 있어서 객관성 있는 군집 결정 방안을 제시하여 준다. 각 개체들이 특정 군집에 속하게 될 퍼지 멤버 함수값을 원소로 하는 유사도 행렬을 통하여 군집화를 수행한다. 본 논문에서는 차원 축소기법의 하나인 주성분 분석과 강력한 통계적 학습 이론인 베이지안 학습을 결합한 군집화 모형을 제안하여, 객관적인 퍼지 군집화를 수행하였다. 제안 알고리즘의 성능 평가를 위하여 UCI Machine Learning Repository의 Iris와 Glass Identification 데이터를 이용한 실험 결과를 제시하였다.

Fuzzy Clustering Model using Principal Components Analysis and Naive Bayesian Classifier

Sung Hae Jun†

ABSTRACT

In data representation, the clustering performs a grouping process which combines given data into some similar clusters. The various similarity measures have been used in many researches. But, the validity of clustering results is subjective and ambiguous, because of difficulty and shortage about objective criterion of clustering. The fuzzy clustering provides a good method for subjective clustering problems. It performs clustering through the similarity matrix which has fuzzy membership value for assigning each object. In this paper, for objective fuzzy clustering, the clustering algorithm which joins principal components analysis as a dimension reduction model with bayesian learning as a statistical learning theory. For performance evaluation of proposed algorithm, Iris and Glass identification data from UCI Machine Learning repository are used. The experimental results shows a happy outcome of proposed model.

키워드 : 주성분 분석(PCA : Principal Components Analysis), 나이브 베이지안 분류기(NBC : Naive Bayesian Classifier), 퍼지 군집화(Fuzzy Clustering)

1. 서 론

전통적인 기계 학습(machine learning) 전략인 군집화(clustering)는 주어진 전체 데이터를 서로 유사한 몇 개의 집단으로 그룹화 한다. 이때 사용되는 유사도 측도(similarity measure)로서는, 주로 거리(distance)에 기반한 측도를 사용한다. 특히 퍼지(Fuzzy) 군집화 전략에서는 유사도 측도로서 각 개체가 특정 군집에 속하게 되는 퍼지 멤버 함수를 사용한다. 즉 모든 개체에 대하여 각 군집에 대한 소속 가능도를 나타내는 유사도 행렬이 구해지고, 이 행렬을 이용하여 최종적인 퍼지 군집화가 수행된다. 퍼지 군집화의 유사도를 나타내는 분할 행렬(partition matrix)의 각 원소는 퍼지 C-평균(Fuzzy C-Means : FCM) 등에서 다양한 방법을 통하여 구해진다[8]. 본 논문에서는 이러한 퍼지 군집화의 유사도 행렬을 구하기 위하여 베이지안 학습(Bayesian learning)의 사후 확률 분포(posterior probability distribution)를 이용한 나이브 베이지안 분류기(Naive Bayesian Classifier : NBC)를 사용하였고, 초기 군집수 결정은 주성분 분석(Principal Components Analysis : PCA)을 통하여 결정된 상위 3개의 보유 주성분들에 대한 3차원 산점도(scatter plot)를 이용하였다. 제안하는 알고리즘의 성능을 확인하기 위한 실험을 위하여 기존의 기계 학습 분야에서 군집화의 성능 평가를

† 정 회 원 : 청주대학교 통계학과 교수
논문접수 : 2003년 11월 7일, 심사완료 : 2004년 5월 13일

위해 많이 사용되고 있는 Fisher의 Iris 데이터와 German의 Glass Identification 데이터를 이용하였다[14].

2. 군집화를 위한 퍼지 시스템 구조

퍼지 군집화에서는 군집화를 위한 유사도 정보를 갖는 분할 행렬 U를 구한다. U의 각 원소인 u_{ik}는 개체 i가 집단 k에 속하게 될 멤버 함수값을 나타낸다[6]. 일반적으로 u_{ik}는 다음의 조건식을 만족한다[10].

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

즉 한 개의 개체에 대하여 모든 가능한 군집에 대한 소속 가능도의 합은 1이 된다. FCM도 퍼지 군집화 기법 중의 하나이다. FCM은 식 (2)의 가중 급내 등급 제곱합(weighted within-class sum of square)을 최소화 하여 군집화를 수행한다[4].

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

위 식에서 $v_k = (v_{ka})$(k=1, ⋯, K, a=1,⋯, p)는 집단 k의 중심값을 나타내고, $x_i = (x_{ia})$(i=1, ⋯, n, a=1,⋯, p)는 i번째 개체를 나타낸다. 그리고 $d^2(x_i, v_k)$는 x_i와 v_k 간의 유클리디안 거리(Euclidean distance)를 나타낸다. m은 1에서 ∞까지의 값을 가지며 군집화의 퍼지화(fuzziness) 정도를 결정한다[12]. 즉 식 (2)를 최소화하는 U와 $v_1, \cdots, v_k$를 결정하여 주어진 학습 데이터를 군집화한다.

3. 퍼지 군집화를 위한 제안 알고리즘

3.1 PCA를 이용한 초기 군집수 결정

3.1.1 주성분의 개념

2보다 크거나 같은 p개 이상의 반응변수들에 대하여 얻어진 다변량 데이터의 처리를 위한 PCA는 입력 변수들 상호간의 상관 관계를 계산하여 다차원 변수들의 차원을 축소하는 통계적 기법이다. 즉, 변수들을 선형 변환(linear transformation)시켜, 주성분이라는 서로 독립적인 새로운 변수들을 유도한다. 이때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도의 순서를 생각할 수 있는데, 그들 중 상위 m개(m<<p)의 주성분에 의하여 원래 자료에 내재하는 전체 변이 중 가능한 한 많은 부분이 보유되도록 변환시킴으로서 정보의 손실을 최소화하는 차원의 축소(dimension reduction)가 이루어진다[1].

서로 상관되어 있는 $p(\geq 2)$개의 확률변수($X_1, X_2, \cdots, X_p$)를 원소로 하는 확률벡터 X의 평균벡터 μ와 공분산행렬 Σ를 다음과 같이 표현한다.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_i \\ \cdot \\ \cdot \\ X_p \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_i \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \sigma_{1j} & \cdot & \cdot & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdot & \cdot & \sigma_{1j} & \cdot & \cdot & \sigma_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{i1} & \sigma_{i2} & \cdot & \cdot & \sigma_{ij} & \cdot & \cdot & \sigma_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdot & \cdot & \sigma_{pj} & \cdot & \cdot & \sigma_{pp} \end{bmatrix} \tag{3}$$

식 (3)에서 $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ji}$는 X_i와 X_j의 공분산(covariance)이다. 그리고 $\mu_i = E[X_i]$는 X_i의 평균이다. σ_{ii}는 i번째 변수의 분산이고 σ_{ij}는 i번째 변수와 j번째 변수의 공분산을 나타낸다. PCA는 입력벡터 X를 선형변환시켜 정보의 손실을 최소화하면서 p보다 매우 작은 m개의 새로운 인공변수를 생성함으로서, p차원 변이를 m차원으로 축소하여 전체 데이터의 특성을 요약하여 전체 변수들간의 복잡한 구조를 파악하고자 하는 것이다. 이 분석은 X의 원소들 간의 상관구조관계를 나타내는 Σ를 분석대상으로 한다.

Σ의 p개의 고유값(eigen value) δ_j들을 크기순으로 배열하고 각각의 고유값에 대응되는 고유벡터(eigen vector), r_j의 짝들을 $(\delta_1, r_1), (\delta_1, r_2), \cdots, (\delta_p, r_p)$라 하고 δ_j들을 크기순으로 배열하면 다음 식과 같이 표현된다.

$$\sum r_j = \delta_j r_j, \; j = 1, 2, \cdots, p \tag{4}$$

식 (4)에서 전체 고유값들은 $\delta_1 \geq \delta_2 \geq \delta_3 \cdots \geq \delta_p$의 관계를 갖게 된다. 이 값에 의해 p개의 입력변수를 m개의 주성분으로 차원 축소하여 데이터의 특성을 파악하게 된다. 이 때 주성분의 개수는 다음 절에서와 같이 몇 가지 판정 기준을 통해 결정하게 된다.

3.1.2 보유 주성분의 수에 관한 판정기준

우선 전체 변이에 대한 공헌 정도로 결정할 수 있다. 보유 주성분들이 전체 분산에 대해 주어진 일정비율 (예를 들어 80~90%) 이상을 설명할 수 있게 하기 위한 최소 필요 개수의 주성분을 보유하게 하는 방법이다. 다음으로 고유값의 크기를 이용할 수 있다. 이 기준은 주성분으로 보유되기 위해서 대응되는 고유값은 적어도 1이상이 되어야 한다는 것으로서, 이는 Kaiser의 규칙(Kaiser, 1960)으로 알려져 있다.

3.1.3 주성분 산점도를 통한 초기 군집수 결정

본 논문에서는 군집화를 위한 학습 데이터의 입력 변수들에 대한 PCA를 수행하여 설명력이 높은 상위 3개의 주성분을 결정하고 이들 주성분들을 각 축으로 하는 3차원 산점도를 그리고 이를 관찰하여 초기 군집수를 시각적으로 결정하였다[3].

3.2 베이지안 학습을 이용한 군집화

본 논문에서는 일반적으로 분류 알고리즘으로 사용되는

NBC를 PCA와 결합하여 최적 군집화가 가능한 모형을 제안하였다. PCA를 통한 초기 군집수를 바탕으로 각 군집에 할당된 개체들의 입력벡터에 대한 평균(mean)과 표준편차(standard deviation : s.d.)를 이용한 사전 확률 분포를 구하고 계속해서 베이즈 정리(Bayes' Theorem)를 적용한 사후 확률 분포를 이용하여 퍼지 군집화를 수행하였다. 제안 모형과 기존의 군집화 모형들과의 비교를 위한 측도로서는 분산과 군집수에 기반한 VC 측도(variance criterion measure)를 사용하였다[2].

3.2.1 나이브 베이지안 분류기

베이지안 분류기(classifier)는 주어진 학습 데이터의 각 개체가 특정 클래스에 속할 확률을 예측해 주는 통계적 분류기이다. 특히 베이지안 분류기는 대용량 데이터로부터 분류 모형을 수행해야 하는 데이터 마이닝 분야에 적용되어 좋은 결과를 제공해 주고 있다[4, 9]. 이는 나이브 베이지안 분류기라고 불려지는 베이지안 분류 모형에서는 각 입력 변수들 간의 독립을 가정하여 사후확률을 계산하기 때문이다. 사후확률은 다음 식과 같은 베이즈 정리에 의해 구해진다.

$$P(C_i) = \frac{P(X|C_i)P(C_i)}{P(X)} \propto P(X|C_i)P(C_i) \quad (5)$$

식 (5)에서 X는 입력벡터, $(x_1, \cdots, x_p)$를 나타내고, C_i는 목표변수의 k개의 클래스, $(C_1, \cdots, C_k)$중에서 i번째 클래스를 나타낸다. 위 식에서 입력벡터의 차원 p가 커지면 속성들 간의 연관성이 존재할 수 있기 때문에 $P(X|C_i)$의 계산 비용이 매우 커지게 된다. NBC에서는 이러한 계산 비용을 최소화 하기 위하여 속성들 간의 조건부 독립성(conditional independence) 가정을 한다. 이 가정은 각 속성의 값들이 다른 속성의 값과 조건 독립으로써 속성들 간에 종속 관계가 존재하지 않는다는 것을 의미한다. 따라서 $P(X|C_i)$는 다음 식과 같이 간단히 계산될 수 있다.

$$P(X|C_i) = \prod_{j=1}^{p} P(x_j|C_i) \quad (6)$$

본 논문에서는 식 (6)의 $P(\cdot)$의 확률 구조를 가우시안 분포(Gaussian distribution)로 하였다. 이는 대부분의 데이터들의 평균이 표본수의 증가에 따라 가우시안 분포로 접근(convergence)하는 중심극한 정리(central limit theorem)에 의하기 때문이다. 하지만 가우시안 접근 분포를 사용하지 않고 감마(gamma), 코쉬(cauchy) 등 특정 분포를 사용할 수도 있다. 이 때에는 사후 확률 계산에 있어서 마코프체인 몬테칼로(markov chain monte carlo : MCMC)와 같은 고급 베이지안 계산(advanced Bayesian computing) 기법을 요구하게 된다[13]. 본 논문에서 사용되는 NBC는 베이즈 정리를 이용하여 개체 X를 다음의 조건을 만족하는 C_i에 할당한다.

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \text{ for } 1 \le j \le k,\ i \ne j \quad (7)$$

즉, X는 k개의 클래스 중에서 가장 큰 사후 확률값을 갖게 되는 집단으로 할당된다.

3.3 베이지안 퍼지 군집화

학습 데이터(training data)의 각 개체가 특정 군집에 속할 퍼지 멤버 함수를 나타내는 퍼지 군집화의 분할 행렬 U의 각 원소는 식 (1)로부터 확률과 같은 구조를 갖게 됨을 알 수 있다. 본 논문에서는 주어진 데이터로부터 나이브 베이지안 학습을 통한 최종 사후 확률 분포의 확률값으로서 퍼지 군집화를 위한 분할 행렬 U를 결정하였다. 퍼지 군집화를 위한 나이브 베이지안 학습에 사용되는 데이터 구조는 다음 식과 같다. 식 (8)은 전체 n개의 데이터 중에서 i번째 데이터에 대한 구조를 나타내고 있다[5].

$$x_1^{(i)}, \cdots, x_N^{(i)} \sim iid\ sample\ from\ \pi_i = N(\theta_i, \Sigma_i) \quad (8)$$

즉 $x_1^{(i)}, \cdots, x_N^{(i)}$는 π_i 분포를 따르는 집단 i로부터 추출된 N_i개의 표본 데이터라고 가정한다. 위 식에서 'i.i.d.(independent, identical distributed) sample'은 임의 표본(random sample)을 의미한다. 또한 π_i는 평균 벡터(mean vector) θ_i와 분산-공분산 행렬(variance-covariance matrix) Σ_i를 갖는 가우시안 분포(Gaussian distribution)라고 가정한다. 식 (8)의 데이터 구조로부터 각 집단의 사전 확률 분포(prior probability distribution)도 역시 가우시안 분포로 결정하였다. 이는 베이지안 학습의 사후 확률 분포의 계산을 쉽게 할 수 있는 공액 분포(conjugate distribution)의 특성을 이용하기 위함이다. 만약 공액 확률 분포를 사용하지 않는다면 확률적 모의 실험을 통하여 사후 확률 분포를 계산해야 하는 MCMC 기법을 사용해야 한다. 이 방법은 매우 많은 계산 비용(computing cost)을 요구한다[7, 11]. 주어진 학습 데이터에 대한 분포인 우도 함수(likelihood function)도 마찬가지 이유로 가우시안 분포로 결정하였다. 따라서 군집화의 최종 U의 원소인 퍼지 멤버함수를 결정하기 위한 사후 확률 분포의 구조도 가우시안 분포가 된다. 다음은 나이브 베이지안 분류기를 통한 퍼지 군집화의 분할 행렬 U의 원소인 퍼지 규칙을 생성하는 알고리즘이다.

[Naive Bayesian Classifier based Fuzzy Rule Extraction algorithm for Clustering]

(단계 1) 분포의 결정

사전 확률 분포의 결정 : $N(\theta_i, \Sigma_i)$

~ Conjugate distribution(Gaussian)

학습 데이터의 우도 함수 결정 : $(x|\pi_i)$

~ Gaussian distribution

(단계 2) 사후 확률 분포의 계산

Bayes' Theorem 이용

Posterior ∝ 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.

$$maxarg_{i \in \{1,2,\cdots,K\}} p(x \in \pi_i | x) \quad (9)$$

위의 식 (9)는 확률 구조이기 때문에 퍼지 군집화의 조건인 식 (1)을 만족하게 된다. 따라서 베이지안 학습을 통하여 퍼지 군집화를 위한 유사도 분할 행렬인 U를 구하였다.

3.4 제안 군집화 알고리즘의 요약

본 논문에서 제안하는 군집 알고리즘은 다음과 같이 3개의 단계를 거친다.

(단계 1) PCA 수행

(1-1) P차원 입력 벡터에 대한 3(or 2)개의 주성분 추출 (3(2 or)<< P)

(1-2) 3차원(or 2차원) 산점도를 통한 초기 군집수 l을 결정

(단계 2) 군집 초기화

(2-1) l개의 군집노드에 대한 입력속성들의 평균($\bar{x}$)과 표준편차(s) 계산

(2-2) 각 군집에 대한 가우시안 분포 $P(G_i)$ 결정 (j = 1, 2, ⋯, l)

$$P(x | G_j) = \frac{1}{\sqrt{2\pi} s_{G_j}} e^{-\frac{(x - \bar{G}_j)^2}{2s_j^2}}$$

$\bar{G}_j$: j번째 군집노드의 평균

$s_{G_j}^2$: j번째 군집노드의 분산

(단계 3) NBC를 이용한 군집화

(3-1) $P(G_i)$의 계산

$$P(G_i) = \frac{N_i}{N}$$

N : 군집으로 선택된 각 노드에 속한 개체들의 총합

N_i : i번째 군집노드의 개체수

(3-2) 사후확률의 계산

$$P(G_i | x) = \frac{P(x | G_i) P(G_i)}{P(x)} \propto P(x | G_i) P(G_i)$$

(3-3) 개체의 군집할당

$$maxarg_{(i=1,2,\cdots,l)} P(G_i | x)$$

위의 3단계의 군집화 과정을 거쳐서 최종적으로 개체 x는 l개의 군집중에서 사후 확률이 가장 큰 군집에 할당되게 된다.

3.5 최종 군집 결과에 대한 평가

결정된 군집수에 대한 성능평가의 기준은 다음의 두 가지 조건을 포함해야 한다.

① 각 군집의 평균 밀도(average density)

② 군집수 증가에 대한 불이익(penalty)

위의 두 조건을 만족하기 위해 군집수에 대한 성능평가의 기준으로 VC 측도를 사용한다. VC 측도는 다음 식과 같이 정의된다.

$$VC_M = \sum_{i=1}^{M} V_i / M + 0.1 * M \quad (10)$$

위의 식에서 M은 군집의 수를, V_i는 i번째 군집의 평균 분산을 의미한다. 앞의 분산의 평균은 각 군집의 평균밀도를 의미하고, 뒤의 군집수는 군집수 증가에 따른 불이익을 의미한다. 따라서 위의 식은 좋은 성능의 군집에 대해서 보다 작은 값을 가지게 된다. 두 항(term)의 균형을 위한 상수는 명확한 군집의 구분을 가지는 인공 데이터를 이용한 휴리스틱(heuristic) 실험을 통해 0.1로 결정되었다.

4. 실험 및 결과

본 논문에서 제안하는 군집화 알고리즘의 성능 평가를 위하여 UCI Machine Learning Repository로부터 Iris와 Glass 데이터를 이용하였다. 이들 데이터는 기존의 많은 기계 학습 알고리즘의 객관적인 성능 평가를 위한 시험 데이터로 사용되었다. 따라서 본 실험에서에도 이들 데이터를 이용하였다.

4.1 Iris 데이터를 통한 성능 평가

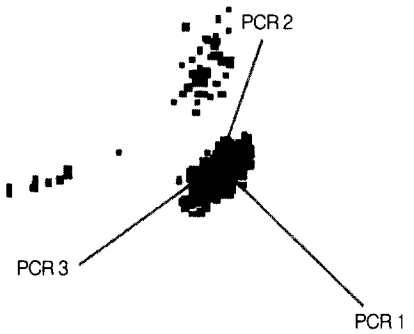
Fisher의 Iris 데이터는 4개의 입력 변수와 1개의 출력 변수로 이루어져 있다. 구체적으로 4개의 입력변수들은 각각 꽃의 외형을 결정하는 값을 가지고 있고, 1개의 출력변수는 꽃의 종류를 나타낸다. 하지만 군집화를 위한 제안 모형의 실험을 위해서는 출력변수는 사용하지 않았다. 총 150개의 학습 개체로 이루어진 Iris 데이터는 입력 변수들에 대한 다음과 같은 기본 통계량을 가지고 있다.

〈표 1〉 Iris 데이터의 입력 변수들의 통계량

변 수	mean	s.d.	min	max
SepalLength	58.4333	8.2807	43	79
SepalWidth	30.5733	4.3587	20	44
PetalLength	37.5800	17.6530	10	69
PetalWidth	11.9933	7.6224	1	25

4.1.1 PCA를 통한 초기 군집수 결정

Iris 데이터의 4개의 입력 변수들에 대한 PCA 결과로 얻어진 3개의 보유 주성분을 이용한 3차원 산점도는 다음과 같다.



(그림 1) 주성분에 의한 초기 군집수 결정

(그림 1)을 보면 전체 데이터가 3개의 군집으로 나뉘어지고 있음을 알 수 있다. 따라서 NBC를 위한 Iris 데이터의 초기 군집수를 3개로 결정하였다. 3개의 군집들의 4개의 입력 변수는 다음과 같은 평균과 분산을 갖는 구조가 된다.

$$G_1 : (\bar{x}_{11}, S^2_{x_{11}}), (\bar{x}_{12}, S^2_{x_{12}}), (\bar{x}_{13}, S^2_{x_{13}}), (\bar{x}_{14}, S^2_{x_{14}})$$
$$G_2 : (\bar{x}_{21}, S^2_{x_{21}}), (\bar{x}_{22}, S^2_{x_{22}}), (\bar{x}_{23}, S^2_{x_{23}}), (\bar{x}_{24}, S^2_{x_{24}})$$
$$G_1 : (\bar{x}_{11}, S^2_{x_{11}}), (\bar{x}_{12}, S^2_{x_{12}}), (\bar{x}_{13}, S^2_{x_{13}}), (\bar{x}_{14}, S^2_{x_{14}})$$

위의 3개의 군집에 대한 각각의 평균과 분산을 이용하여 사후확률 분포로 사용할 가우시안 분포의 모수를 결정한다. 가우시안 분포의 모수는 주어진 데이터의 평균과 분산이기 때문이다. 따라서 최종 군집화를 위한 NBC의 사후 확률분포는 다음 식과 같이 구한다.

$$P(x|G_j)P(G_j) = \prod_{i=1}^{4} P(x_i|G_j)P(G_j) \quad (j=1,2,3) \quad (11)$$

결론적으로 식 (11)을 통하여 특정 입력에 대한 3개의 군집에 대한 사후 확률 $P(G_1|\underline{x})$, $P(G_2|\underline{x})$, 그리고 $P(G_3|\underline{x})$를 구하게 된다. 그리고 이들 3개의 사후 확률 값을 보고 $\underline{x}$를 가장 큰 사후 확률값을 갖는 군집에 할당하게 된다.

다음 표는 3개의 군집에 대한 Iris 데이터의 4개의 입력 변수들에 대한 평균과 분산이다. 이 값들을 이용하여 사전 분포로 사용되는 가우시안의 모수를 결정한다.

〈표 2〉 각 군집의 입력벡터에 대한 가우시안 모수값

Group	SepalLength		SepalWidth		PetalLength		PetalWidth	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
G1	48.71	3.13	30.60	3.68	14.10	1.84	2.19	0.98
G2	61.34	5.17	27.79	2.99	41.47	5.12	12.98	2.01
G3	66.18	6.65	31.43	3.97	54.39	5.68	20.01	2.81

제안 군집화 기법인 PCA-NBC에 의한 군집화 결과의 VC 값이 기존의 군집 분석 기법과 비교한 결과가 다음의 표에 나타났다.

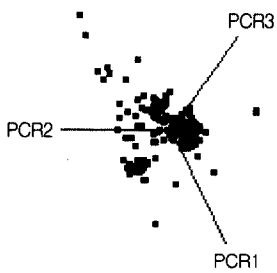
〈표 3〉 제안 모형과 기존의 모형의 VC 값

군집 기법	VC 값
PCA-NBC	0.58
Batch SOM	0.97
K-Means	0.73
Hierarchical clustering	0.84
FCM	0.68

위의 표에 의하면 본 논문에서 제안한 군집화 기법의 VC 값이 가장 작음을 알 수 있다. 이는 제안 방법에 의한 군집화 결과가 다른 군집화 기법에 비해 군집내 분산은 상대적으로 작고, 군집간 분산은 크게 나타나고 있음을 확인할 수 있었다. 모형 평가의 공정성을 유지하기 위하여 제안 방법과 비교되는 위의 다른 군집화 기법의 초기 군집수도 모두 3으로 통일하였다.

4.2 Glass Identification 데이터를 이용한 성능 평가

German의 Glass 데이터는 총 214개의 학습 자료로 이루어져 있다. 입력 변수들은 1개의 굴절률 변수와 성분을 나타내는 8개의 변수들(나트륨, 마그네슘, 알루미늄, 실리콘, 칼륨, 칼슘, 바륨, 철)로 이루어져 있다. 제안 알고리즘의 1단계 PCA 수행 결과 다음 그림과 같은 3차원 주성분 산점도를 얻었다.



(그림 2) Glass 데이터의 주성분 산점도

위 그림을 통하여 초기 군집수는 5개로 결정하였다. 5개의 군집에 대한 평균과 표준 편차를 구하고 이를 이용하여 나이브 베이지안 분류기를 통한 최종 군집 결과의 VC 값은 다

음과 같다.

〈표 4〉 제안 모형과 기존의 모형의 VC 값

군집 기법	VC 값
PCA-NBC	0.36
Batch SOM	0.89
K-Means	0.66
Hierarchical clustering	0.73
FCM	0.46

Glass 데이터를 이용한 군집화 결과에서도 제안 방법에 의한 군집화 결과가 가장 유사성이 높은 개체들끼리 묶기고 있음을 알 수 있었다.

5. 결론 및 향후 연구과제

본 논문에서는 통계적 차원 축소 기법인 PCA와 베이지안 학습에 기반한 분류 모형인 NBC를 이용하여 최적의 군집화를 위한 군집화 알고리즘을 제안하였다. 기존의 군집화 알고리즘은 단일 군집화 과정을 거치지만 제안 기법은 초기 군집수 결정과 최적 군집 할당을 위한 퍼지 멤버함수 계산과정의 2단계 군집화 전략을 취하고 있다.

전통적으로 기계 학습 알고리즘들의 성능 평가를 위하여 사용되었던 2개의 학습 데이터를 이용한 실험에서 제안 기법의 성능이 기존의 것들에 비해 우수하게 나타나고 있음을 확인하였다. 향후 MCMC에 의한 베이지안 분류기를 이용하여 더 좋은 군집화 결과를 얻을 수 있는 연구가 기대되어 진다.

참 고 문 헌

[1] 김기영, 전명식, 다변량 통계자료 분석, 자유아카데미, 1994.

[2] 박민재, 전성해, 오경환, "붓스트랩 기법과 유전자 알고리즘을 이용한 최적 군집수 결정", 퍼지및지능시스템학회논문지, 제13권 제1호, pp.12-17, 2003.

[3] 전성해, 오경환, "MCMC 결측치 대체와 주성분 산점도 기반의 SOM을 이용한 희소한 웹 데이터 분석", 정보처리학회논문지D, Vol.10-D, No.2, pp.277-282, April, 2003.

[4] 한진우, 전성해, 오경환, "군집화를 위한 베이지안 학습 기반의 퍼지 규칙 추출", 한국정보과학회 2003 춘계학술발표논문집(II), April, 2003.

[5] J. S. Liu, J. L. Zhang, M. L. Palumbo, C. E. Lawrence, "Bayesian Clustering with Variable and Transformation Selections," Bayesian Statistics 7, Oxford University Press, 2003.

[6] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum Press, 1987.

[7] C. M. Bishop, "Neural Networks for Pattern Recognition," Clarendon Press : Oxford, 1998.

[8] D. Dumitrescu, B. Lazzerini, L. C. Jain, "Fuzzy Sets and their application to Clustering and Training," The CRC Press, 2000.

[9] J. Han, M. Kamber, "Data Mining : Concepts and Techniques," Morgan Kaufmann, 2000.

[10] R. J. Hathaway, J. C. Bezdek, "Switching Regression Models and Fuzzy Clustering," IEEE Trans. Fuzzy Sets, Vol.1, pp.195-204, 1993.

[11] S. J. Press, "Bayesian Statistics : Principles, Models, and Applications," John Wiley & Sons, 1989.

[12] H. J. Zimmermann, "Fuzzy Set Theory and Its Application," Kluwer Academic Publishers Group, 2001.

[13] C. P. Robert, G. Casella, "Monte Carlo Statistical Methods," Springer, 1999.

[14] http://www.ics.uci.edu/~mlearn/.

전 성 해

e-mail : shjun@chongju.ac.kr

1993년 인하대학교 통계학과(학사)

1996년 인하대학교 대학원 통계학과 (이학석사)

2001년 인하대학교 대학원 통계학과 (이학박사)

1996년~1997년 효성그룹 전자통신연구소 연구원

2000년~2001년 NCR Korea 데이터마이닝 컨설턴트

2002년~2003년 경기대학교 정보통신대학원 겸임교수

2001년~현재 서강대학교 대학원 컴퓨터학과 공학박사수료

2003년~현재 청주대학교 통계학과 전임강사

관심분야 : 유비쿼터스, 데이터공학, 생물정보학, 기계학습