군집화를 위한 베이지안 학습 기반의 퍼지 규칙 추출

한진우 전성해[0] 오경환
서강대학교 컴퓨터학과, 청주대학교 통계학과[0]
{hey_han@ailab, shjun@ailab[0], kwoh@ccs}.sogang.ac.kr

Bayesian Learning based Fuzzy Rule Extraction for Clustering

Jin-Woo Han Sung-Hae Jun[0] Kyung-Whan Oh
Dept. of Computer Science, Sogang University
[0]Dept. of Statistics, Chongju University

요 약

컴퓨터 학습의 군집화는 주어진 데이터를 서로 유사한 몇 개의 집단으로 묶는 작업을 수행한다. 군집화를 위한 유사도 결정을 위한 측도는 많은 기법들에서 매우 다양한 측도들이 사용되고 또한 연구되어 왔다. 하지만 군집화의 결과에 대한 성능측정에 대한 객관적인 기준 설정이 어렵기 때문에 군집화 결과에 대한 해석은 매우 주관적이고 애매한 경우가 많다. 퍼지 군집화는 이러한 애매한 군집화 문제에 있어서 융통성 있는 군집 결정 방안을 제시해 준다. 각 개체들이 특정 군집에 속하게 될 퍼지 멤버 함수값을 원소로 하는 유사도 행렬을 통하여 군집화를 수행한다. 본 논문에서는 베이지안 학습을 통하여 군집화를 위한 퍼지 멤버 함수값을 구하였다. 본 연구에서는 최적의 퍼지 군집화 수행을 위하여 베이지안 학습 기반의 퍼지 규칙을 추출하였다. 인공적으로 만든 데이터와 기존의 기계 학습 데이터를 이용한 실험을 통하여 제안 방법의 성능을 확인하였다.

1. 서 론

전통적인 기계 학습(machine learning) 기법인 군집화(clustering)는 주어진 전체 데이터를 서로 유사한 몇 개의 집단으로 그룹화한다. 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 분할 행렬을 이용하여 퍼지 군집화가 수행된다.

퍼지 군집화의 유사도를 나타내는 분할 행렬의 개개의 원소는 퍼지 C-평균(Fuzzy C-Means: FCM) 등에서 다양한 방법을 통하여 구해진다. 본 논문에서는 이러한 퍼지 군집화의 유사도 행렬을 베이지안 학습(Bayesian learning)의 사후 확률 분포(posterior probability distribution)를 이용하여 구하고 이를 통하여 주어진 학습 데이터를 군집화 하였다.

제안하는 알고리즘의 실험을 위하여 기존의 기계 학습 데이터인 Fisher의 Iris데이터와 SAS E-Miner를 통하여 인공적으로 생성한 데이터를 이용하였다[8][9].

2. 군집화를 위한 퍼지 시스템 구조

퍼지 군집화에서는 군집화를 위한 유사도 정보를 가지는 분할 행렬 U를 구한다. U의 각 원소인 u_{ik}는 개체 i가 집단 k에 속하게 될 멤버 함수값을 나타낸다[2]. 일반적으로 u_{ik}는 다음의 조건식을 만족한다.

$$u_{ik} \in [0,1], \quad \sum_{i=1}^{K} u_{ik} = 1 \qquad \text{식 (1)}$$

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 급내의 등급 제곱합(weighted within-class sum of square)을 최소화 하여 군집화를 수행한다[4].

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

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_k간의 유클리디안 거리(Euclidean distance)를 나타낸다. m은 1에서 ∞까지의 값을 가지며 군집화의 퍼지화(fuzziness) 정도를 결정한다[6].

즉 식 (2)를 최소화하는 U와 $v_1, \ldots, v_k$를 결정하여 주어진 학습 데이터를 군집화한다.

3. 베이지안 학습을 통한 군집화 퍼지 규칙의 추출

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 베이지안 학습을 통하여 최종 사후 확률 분포로서 퍼지 군집화를 위한 분할 행렬 U를 결정하였다.

퍼지 군집화를 위한 베이지안 학습에 사용되는 데이터 구조는 다음 식과 같다. 식 (3)은 전체 n개의 데이터 중에서 i번째 데이터에 대한 구조를 나타내고 있다[1].

$$x_1^{(i)}, \ldots, x_{N_i}^{(i)} \sim iid \ \ sample \ \ from \ \ \pi_i = N(\theta_i, \Sigma_i) \quad \text{식 (3)}$$

즉 $x_1^{(i)}, \ldots, x_{N_i}^{(i)}$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 (3)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 모의실험을 통하여 사후 확률 분포를 계산해야 하는 마코프 체인 몬테 칼로(Markov Chain Monte Carlo: MCMC) 기법을 사용해야 한다. 이 방법은 매우 많은 계산 비용(computing cost)을 요구한다[3][5].

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 베이지안 학습을 통한 퍼지 군집화의 분할 행렬 U의 원소인 퍼지 규칙을 생성하는 알고리즘이다.

// Bayesian Learning based Fuzzy Rule Extraction algorithm: Clustering approach //

(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

(2) 사후 확률 분포의 계산
Bayes' Theorem 이용
Posterior $\propto$ 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

(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.

$$maxarg_{i \in \{1,2,\ldots,K\}} p(x \in \pi_i | x) \quad \text{식 (4)}$$

위의 식 (4)는 확률 구조이기 때문에 퍼지 군집화의 조건인 식 (1)을 만족하게 된다. 따라서 베이지안 학습을 통하여 퍼지 군집화를 위한 유사도 분할 행렬인 U를 구하였다.

다음 그림은 제안하는 알고리즘에 의해 개체가 군집에 할당되는 과정을 도식화하였다.

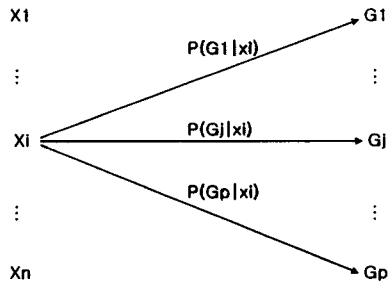

그림 1. 사후 확률에 의한 학습 데이터의 군집화

그림 1에 의하면 개체 Xi는 다음의 식을 만족하는 집단 G^*에 할당한다.

$$P(G^* | x_i) = max_{j \in \{1,2,\ldots,p\}} P(G_j | x_i) \quad \text{식 (5)}$$

각 개체에 대한 각 군집의 사후 확률값을 계산하여 가장 큰 값을 갖는 군집에 해당 개체를 할당하는 식 (5)의 군집 판정 기준을 이용하여 퍼지 군집화를 수행한다.

4. 실험 및 결과

본 논문의 제안 알고리즘의 실험을 위하여 2개의 기계학습 데이터를 사용하였다. 5개의 집단을 갖는 인공 데이터는 데이터 마이닝 툴인 SAS사의 Enterprise Miner를 이용하였다. 또한 UCI Machine Learning Repository 로부터 3개의 군집을 이루고 있는 Fisher의 Iris 데이터를 이용하였다. 다음의 표 1은 인공 데이터를 나타내고 있다.

표 1. 5개의 군집을 갖는 인공 데이터

	y1	y2	⋯	y10
x1	2	1	⋯	9
x2	6	5	⋯	3
⋮	⋮	⋮	⋱	⋮
x100	11	6	⋯	1

표 1에 의하면 인공 데이터는 100개의 개체(x1,...,x100)로 이루어지고 각 개체는 10개의 군집 변수(y1,...,y10)를 갖는다. 주어진 학습 데이터에 대하여 본 논문에서 제안하는 베이지안 학습 기반의 퍼지 군집화를 위한 사후 확률의 계산 결과는 다음 표에 나타나 있다.

표 2. 5개의 군집에 대한 각 개체의 사후 확률

	G1	G2	G3	G4	G5
x1	P(G1|x1) =0.31	P(G2|x1) =0.14	P(G3|x1) =0.13	P(G4|x1) =0.19	P(G5|x1) =0.23
x2	P(G1|x2) =0.29	P(G2|x2) =0.09	P(G3|x2) =0.37	P(G4|x2) =0.17	P(G5|x2) =0.08
⋮	⋮	⋮	⋮	⋮	⋮
x100	P(G1|x100) =0.14	P(G2|x100) =0.37	P(G3|x100) =0.11	P(G4|x100) =0.13	P(G5|x100) =0.25

위의 표 2로부터 다음과 같은 퍼지 군집화의 분할 행렬 U를 구축할 수가 있게 된다.

$$U = \begin{pmatrix} 0.31 & 0.14 & 0.13 & 0.19 & 0.23 \\ 0.29 & 0.09 & 0.37 & 0.17 & 0.08 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.14 & 0.37 & 0.11 & 0.13 & 0.25 \end{pmatrix}$$

위의 U 행렬에서 행(row)은 개체이고 열(column)은 군집을 나타낸다. 즉 개체 1은 U의 원소 중에서 가장 큰 퍼지 군집화 멤버 함수값(0.31)을 갖는 군집 1에 할당된다. 결국 베이지안 사후 확률에 의해 구해진 원소들로 이루어진 U는 퍼지 군집화를 위한 유사도 행렬(similarity matrix)이 된다.

다음은 Iris 데이터를 이용하여 퍼지 군집화의 유사도 행렬을 구하였다. Iris 데이터는 총 150개의 개체 수를 가지고 있다. 또한 꽃의 외형을 결정하는 4개의 입력 변수와 3개의 꽃의 종류를 표시하는 한개의 목표 변수로 이루어져 있다. 표 3은 제안 알고리즘에 의한 Iris 데이터의 군집화 결과의 일부이다.

표 3. Iris 데이터에 대한 각 개체의 사후 확률

	G1	G2	G3
x1	P(G1|x1) =0.43	P(G2|x1) =0.27	P(G3|x1) =0.30
x2	P(G1|x2) =0.36	P(G2|x2) =0.18	P(G3|x2) =0.46
⋮	⋮	⋮	⋮
x150	P(G1|x150) =0.09	P(G2|x150) =0.24	P(G3|x150) =0.67

표 2의 결과에서 x1 개체는 가장 사후 확률값이 큰 집단(G1)에 군집화 된다. 이러한 사후 확률을 통하여 전체 데이터의 개개의 개체를 각 군집으로 할당하게 된다. 다음은 Iris 데이터의 군집화를 위한 최종 퍼지 군집화의 분할 행렬 U를 나타내고 있다.

$$U = \begin{pmatrix} 0.43 & 0.27 & 0.30 \\ 0.36 & 0.18 & 0.46 \\ \vdots & \vdots & \vdots \\ 0.09 & 0.24 & 0.67 \end{pmatrix}$$

5. 결론 및 향후 연구과제

본 논문에서는 퍼지 군집화의 분할 행렬에 대한 멤버 함수값의 결정을 위하여 베이지안 학습을 이용하였다. 기존의 FCM 등에서의 상사도 행렬에 비해 학습 데이터의 사전 확률 분포와 데이터에 의한 우도 함수를 사용하여 좀 더 객관적으로 개체를 군집에 할당할 수 있게 되었다. 제안한 베이지안 학습에 의한 사후 확률 분포의 계산은 공액 확률 분포만을 사용하여 모형을 단순화시켰지만 향후에는 가우시안 분포와 같이 분포에 대한 사전 가정이 없이 사후 확률 분포를 계산할 수 있는 MCMC 기법을 사용할 수 있을 것이다. 이때 문제가 되는 것은 계산시간이다. 이러한 계산 시간의 단축 전략은 사후 확률 분포로 부터의 표본 추출 과정을 좀더 빠르고 정확하게 할 수 있는 'Hybrid Monte Carlo simulation' 혹은 'Slice Sampling' 같은 기법을 적용할 수 있다. 이는 향후 연구 과제로 남긴다.

감사의 글

본 연구는 과학 기술부 주관 뇌신경 정보학 사업에 의해 지원되었음.

참 고 문 헌

[1] 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

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

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

[4] R. J. Hathaway, J. C. Bezdek 'Switching Regression Models and Fuzzy Clustering,' IEEE Trans. Fuzzy Syst., 1, 3, 195-204, 1993.

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

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

[7] C. P. Robert, G. Casella, 'Monte Carlo Statistical Methods,' Springer, 1999.

[8] http://www.sas.com

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