한국컴퓨터정보학회 하계학술대회 논문집 제21권 제2호 (2013. 7)

클러스터 내 분별 오류 최소화를 위한 퍼지 클러스터링

허경용[O], 이수종[**]
[*]동의대학교 전자공학과
[**]협성대학교 컴퓨터공학과
e-mail: hgycap@deu.ac.kr[*], sjlee@uhs.ac.kr[**]

Within-Cluster-Discriminative Fuzzy Clustering

Gyeongyong Heo[O], Soojong Lee[**]
[*]Dept. of Electronic Engineering, Dong-Eui University
[**]Dept. of Computer Engineering, Hyupsung University

● 요 약 ●

퍼지 클러스터링은 유사도가 높은 데이터 포인트들이 동일한 클러스터에 포함되도록 하는 대표적인 비교사 학습 방법 중 하나이다. 이 논문에서는 클러스터링을 분류기의 전처리 단계에서 활용할 수 있도록 클러스터 내에서 분류 오류가 최소가 될 수 있도록 클러스터를 생성할 수 있는 새로운 퍼지 클러스터링 방법을 제안한다. 제안하는 클러스터링은 특징 벡터와 함께 클래스 라벨을 활용하므로 분류기와 결합하여 사용할 경우 기존 분류기와 함께 사용할 경우 보다 우수한 성능을 기대할 수 있다.

키워드: 퍼지 클러스터링(fuzzy clustering), 교사 학습(supervised learning), 최소 오류 클러스터링(minimum error clustering)

I. 서 론

컴퍼지 클러스터링은 60년대 Zadeh의 퍼지 입합 이론에서 시작하여 다양한 분야에 적용되고 왔으며 그 중 대표적인 분야가 클러스러팅이다[1]. 퍼지 클러스터링은 대표적인 비교사 학습 방법의 하나로 유사한 속성을 갖는 데이터를 클러스터로 구별하는 전통적인 방법은 물론 융합을 위한 문맥 분할 방법으로도 사용되었다[2]. 하지만 클러스터링은 비교사 학습 방법이며 융합은 교사 학습 방법이므로 융합의 전처리로 사용된 클러스터링이 융합의 측면에서 최적의 결과를 보장하지는 못한다.

이 논문에서는 분류의 관점에서 최적의 클러스터링을 수행하는 새로운 클러스터링 기법, 특히 융합을 위한 문맥 분할의 방법으로서의 클러스터링 기법을 제안한다. 제안하는 클러스터링 기법은 특징공간을 분할함에 있어 각 부분공간에서 국부적으로 오류가 최소화되는 방식으로 클러스터링을 수행함으로써 융합 결과에서도 오류가 줄어들도록 할 수 있다는 아이디어에서 출발한다. 이를 위해 제안한 방법에서는 선형 판별 분석[3]에서와 유사하게 같은 클래스에 속하는 두 데이터 포인트 사이의 거리를 짧고 다른 클래스에 속하는 두 데이터 포인트 사이의 거리를 길다는 제약 조건을 퍼지 클러스터링의 목적함수에 추가함으로써 분류 오류를 최소화하고자 하였으며 실험 결과를 통해 제안한 방법이 효과적임을 확인할 수 있었다.

II. 본 론

N개의 D차원 데이터 $X=\{x_i|1 \leq i \leq N, x_i \in R^D\}$가 주어졌을 때 이를 K개의 클러스터링하기 위해 Fuzzy C-Means(FCM)는 식 (1)의 목적 함수를 최소화 한다.

$$J_{FCM}=\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ki}^{m}\|x_i-v_k\|_A^2=\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ki}^{m}d_{ki}^2 \quad (1)$$

이 때 u_{ki}는 x_i가 k번째 클러스터에 소속되는 정도를 나타내는 소속도 값(membership value)을, v_k는 k번째 클러스터의 중심을, m은 퍼지화 상수(fuzzifier constant)를 나타낸다. 제안하는 클러스터 내 분별 오류를 최소화하는 퍼지 클러스터링은 식 (1)에 클래스 라벨 $Y=\{y_i|1 \leq i \leq N, y_i \in R\}$를 함께 사용하여 동일한 클러스터에 속하는 두 점이 동일한 클래스에 속하는 경우에는 가까이 존재하고, 서로 다른 클래스에 속하는 경우에는 멀리 존재하도록 하는 제약 조건을 추가하여 목적 함수를 식 (2)와 같이 구성하였다.

$$J = \sum_{k=1}^{K}\sum_{i=1}^{N} u_{ki}^{m}\|x_i - v_k\|_A^2 \quad (2)$$
$$+ \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ki}^{m}u_{kj}^{m}d_{\alpha,ij}I(y_i = y_j)$$
$$+ \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ki}^{m}u_{kj}^{m}d_{\beta,ij}I(y_i \neq y_j)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{N} u_{ki}^{m}d_{ki}^2 + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} u_{ki}^{m}u_{kj}^{m}\left(d_{\alpha,ij}^2 I_{ij} + d_{\beta,ij}^2 \overline{I_{ij}}\right)$$

식 (2)에서 $I(\cdot)$는 매개 변수가 참이면 1의 값을, 거짓이면 0의 값을 갖는 지시 함수(indicator function)로 $I(y_i = y_j) = I_{ij}$, $I(y_i \neq y_j) = \overline{I_{ij}}$로 나타내었으며 $d_{\alpha,ij}^2$는 두 점 x_i와 x_j 사이의 거리에 비례하는 함수이며 $d_{\beta,ij}^2$는 두 점 사이의 거리에 반비례하는 함수로 동일한 점에서는 0의 값이 나오는 시그모이드 함수의 변형을 사용하였으며 지시 함수 I_{ij}와 $\overline{I_{ij}}$와 함께 사용되어 클러스터 내의 분리도를 높인다. 식 (2)를 최적화하는 갱신식(updae equation)은 식 (2)의 라그랑지 방정식을 통해 얻을 수 있다.

III. 본 론

실험에 사용한 데이터는 그림 1과 같이 두 개의 클래스에서 생성된 4개의 클러스트 구조를 이루는 데이터이다.

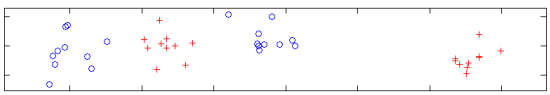

그림 1. 테스트 데이터
Fig. 1. Test data

그림 1을 FCM으로 클러스터링 하는 경우 결과는 그림 2와 같다. 이는 거리가 가까운 점들로 클러스터를 형성하기 때문이다.

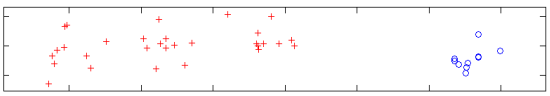

그림 2. FCM으로 클러스터링한 결과
Fig. 2. Clustering result with FCM

그림 2의 경우 왼쪽 클러스터는 서로 다른 클래스에서 생성된 데이터들이 호재되어 있어 비선형의 분류기를 사용하여야만 분류가 가능한 단점이 있다. 이에 비해 제안한 방법으로 클러스터링 한 결과는 그림 3과 같다.

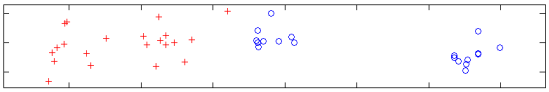

그림 3. 제안한 방법으로 클러스터링 한 결과
Fig. 3. Clustering result with the proposed method

그림 3은 왼쪽과 오른쪽 클러스터 모두 간단한 선형 분류기로 분류가 가능하며 이처럼 단순한 분류기를 사용하므로 잡음 민감도를 줄일 수 있고 분류기의 신뢰도를 높일 수 있다.

IV. 결 론

이 논문에서는 기존의 비교사 학습 방법으로서의 클러스터링이 아닌 분류의 관점에서 최적의 결과를 얻을 수 있는 새로운 클러스터링 방법을 제안하였다. 실험 결과에 나타난 바와 같이 제안한 방법은 분류에 효과적으로 사용될 수 있을 것으로 생각되지만 거리 척도 및 거리 파라미터에 민감하게 반응하는 단점이 있어 현재 신뢰성 있는 거리 척도에 관해 연구 중에 있다.

참고문헌

[1] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, Vol.16, No.3, pp. 645-678, March 2005.

[2] G. Heo, P. Gader and H. Frigui, “A Noise Robust Variant of Context Extraction for Local Fusion,” Proceedings of 2010 IEEE International Conference on Fuzzy Systems, pp. 1-6, July 2010.

[3] A. M. Martinez and A. C. Kak, "PCA versus LDA," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23 No. 2, pp. 228-233, February 2001.