# 불완전 데이터 처리를 위한 퍼지 분류 알고리즘

이찬희* · 박충식** · 우영운***

*동의대학교 디지털미디어공학과

**영동대학교 컴퓨터공학과

***동의대학교 멀티미디어공학과

# Fuzzy Classification Algorithm for Incomplete Data

Chan-Hee Lee* · Choong-shik Park** · Young Woon Woo***

*Dept. of DigitalMedia Eng., Dong-Eui University

**Dept. of Computer Eng., Youngdong University

***Dept. of Multimedia Eng., Dong-Eui University

E-mail : chany1026@deu.ac.kr

## 요 약

패턴 분류 문제는 기계 학습 분야에서 매우 중요한 연구 주제이다. 하지만 불완전 데이터는 실생활에서 매우 빈번히 발생 할 뿐만 아니라 분류 모델의 학습도가 낮다는 문제점을 지니고 있다. 불완전한 데이터를 다루는 것에 대한 많은 방법들이 제안되어 왔지만 대부분의 방법들이 훈련 단계에 집중하고 있다. 본 논문에서는 삼각 형태의 퍼지 함수를 이용하여 불완전 데이터의 분류 알고리즘을 제안한다. 제안한 기법에서는 불완전한 특징 벡터에서의 불완전 데이터를 추론하고 학습하였으며, 추론된 데이터의 가중치를 삼각 퍼지 함수 분류기에 적용하였다. 실험을 통하여 제안한 기법이 상대적으로 높은 인식률을 나타냄을 확인할 수 있었다.

## 키워드

퍼지 분류기, 불완전 데이터, 삼각 퍼지 함수

## Ⅰ. 서 론

패턴 인식 (Pattern Recognition)을 여러 방식으로 정의 할 수 있겠지만 “계산이 가능한 기계적인 장치(컴퓨터)가 어떠한 대상을 인식하는 문제를 다루는 인공지능의 한 분야”라고 정의한다. 정보화 시대에 쉽게 얻을 수 있는 수많은 데이터들을 적적히 표현하는 것은 물론, 다양하고 방대한 자료를 분류하고 처리하기 위한 기술이 다양하게 발전 되어 오고 있다. 그중에서도 인식 기술은 정보를 검색하고 분류, 처리하기 위한 가장 뛰어난 기술이며, 인식 기술의 대표적인 것이 패턴 인식이다. 대표적인 응용 분야로는 문자인식, 생체인식과 인간 행동 패턴 분석, 의료 영상 분석(medical diagnosis) 및 진단 시스템, 도면 인식, 예측 시스템, 보안과 군사 분야 등이 있다.

정보화 시대에의 많은 양의 데이터를 효과적으로 분류하는 작업은 매우 중요할 뿐만 아니라 필수적이다. 하지만 실생활에서 얻어지는 데이터들은 상황에 따라 분실되어지고 외곡 되어 진다. 이러한 불완전 데이터들은 분류 모델의 성능을 저하 시킬 뿐만 아니라, 매우 어려운 문제로 인식된다.

불완전 데이터를 다루기 위한 많은 기법들이 제안되고 있는데, 현재까지의 연구는 베이지안 분류기(Bayesian model)를 이용한 기법[1]과 SVM을 이용한 기법[2~4]이 있고, J. Ross Quinlan[5]이 불완전한 데이터를 무시하고 분류하는 기법을 제안하였으며, 이와 반대로 불완전 데이터를 추론한 후 Fuzzy C-Means를 이용한 클러스터링 기법[6]이 제안되었다.

본 논문에서는 기존의 비지도(unsupervised) 학습인 불완전한 데이터를 이용한 Fuzzy C-Means의 클러스터링(clustering) 기법[6]을 지도 학습(supervised learning)으로써의 퍼지 분류기에 알맞게 개선하여 패턴 분류(classification)에 적용하였다.

제안한 기법들을 이용하여 UCI machine learning repository사이트에서 제공되는 표준 데

이터인 Breast Cancer Wisconsin 데이터와 Ecoli 데이터 세트를 이용하여 실험하고 그 결과를 비교, 분석 하였다.

## II. 제안한 퍼지 알고리즘 기법

### 2.1 분실 데이터의 대체

본 논문에서는 불완전 데이터의 분실 데이터를 추측하고, 복원하여 퍼지 알고리즘에 적용한다. 복원 방법으로서는 각 클래스(class)의 중심을 이용하는 방법과 기존의 방법인 전체 데이터 사이의 거리를 유클리드 거리(Euclidean distance)를 이용하여 유사한 정도를 산출하여 가장 유사한 데이터의 정보로 대체시켜주는 방법을 비교 실험하였다.

불완전 데이터의 간단한 예로 $(1,2,3,4)$ 일 때, $(?,2,3,4)$는 25%의 분실 데이터이며, $(?,2,?,4)$는 50%의 분실 데이터라고 할 수 있다. 본 논문에서 제안하는 기법의 설명에 앞서 설명의 명료함을 돕기 위해 몇 가지 중요 요소들을 제시한다 [6].

정의 1 : 불완전 데이터 $X_A$

$$X_A = \{x_1, x_2, ..., x_n\} \subset R^n$$
$$x_k = \{x_{k1}, x_{k2}, ..., x_{ks}\} \subset R^s$$
$$X_W \subset X_A \text{ and } X_P \subset X_A$$
$$where$$
$$X_W = \{x_k | x_k \in X_A, x_k is a whole datum\}$$
$$X_P = \{x_i | x_j \in X_A, x_i is an \in complete datum\}$$
$$X_M = \{x_{kj} | x_{kj} = ?, 1 \le j \le s, 1 \le k \le n\}$$
$$X_U = \{x_{kj} | x_{kj} = certain value, 1 \le j \le s, 1 \le k \le n\}$$

$$X_W \cap X_P = \phi, X_W \cup X_P = X_A$$
$$X_M \cap X_U = \phi, |X_M| + |X_U| = |X_A|$$

정의 2 : The effect factor $\alpha_{kj}$

$$\alpha_{kj} = \begin{cases} 1 & x_{kj} \in X_U \\ \dfrac{|x_k| - |x_{kj}|}{|x_k||x_{kj}|} & x_k \in X_P, x_{kj} \in X_M \end{cases}$$

정의 3 : The resembling factor $\beta_{ij}$

$$\beta_{ij} = \|x_{ik} - x_{jk}\|, \ (1 \le k \le s) \wedge (x_i \in X_U) \wedge (x_j \in X_M)$$

### 2.2 기존의 처리 기법

기존의 클러스터링 기법에[6]서는 전체 데이터를 대상으로 분실 벡터와 완전한 벡터 간의 닮음 정도를 (정의 3)의 방법으로 계산하여 분실 데이터 값을 가장 닮은 특징 벡터의 특징 값으로 대체하였다. 데이터의 대체 시 (정의 2)의 $\alpha_{kj}$를 계산하여 Fuzzy C-Means 알고리즘에서의 클래스 중심을 $\alpha_{kj}$와 클래스 소속 벡터간의 가중치 평균으로 구한다.

본 논문에서는 기존의 기법을 패턴 분류(classification)에 알맞게 적용하였고, 분실 데이터의 대체 방법을 다음과 같이 개선하여 비교 실험하였다.

### 2.3 제안한 첫 번째 기법

제안한 첫 번째 기법은 다음과 같다.

단계 1 : 각 클래스의 중심을 계산한다.

$$c_k = \{c_{k1}, c_{k2}, ...c_{kn}\} where \quad c_{ki} = \frac{\sum_{j=1}^{\eta_{ki}} x_{kij}}{\eta_{ki}} \qquad (1)$$
$$where \ x_{kij} \in X_U$$

식 (1)의 $c_k$는 각 클래스의 중심이며, 특징별 중심 위치는 분실 데이터를 제외하고 계산한다. $\eta_{ki}$는 $k$번째 클래스의 $i$번째 특징 중 분실되지 않은 데이터의 수를 의미 한다.

단계 2 : 임의의 $c_k$와 분실 데이터 사이의 $\beta_k'$를 식 (2)와 같은 방법으로 구한다.

$$\beta_k' = \|c_{kj} - x_j\|, \ 1 < k \le cn \qquad (2)$$

식 (2)에서의 $cn$은 클래스의 수를 의미하며, $x_j$는 불완전 특징벡터 $x$의 데이터 값 중 분실 되지 않은 데이터 값을 의미한다.

단계 3 : 가장 작은 거리의 $k$의 값에 해당하는 중심 값으로 불완전 데이터 값을 식 (3)과 같이 대체한다.

$$Replace \ x_j \ by \ c_{kj} \ , \ c_{kj} \in c_k \qquad (3)$$

단계 4 : (정의 2)의 $\alpha_{kj}$ 값을 계산한다.

### 2.4 제안한 두 번째 기법

제안한 또 다른 기법은 기존의 방법과 같이 전체 데이터와의 유사성을 계산한 후 인접한 데이터들의 평균값을 대체시키는 방법이다. 자세한 처리 과정은 다음과 같다.

단계 1 : $i$번째 불완전 특징 벡터와 모든 특징 벡터에서의 완전한 벡터간의 거리 집합 $D_i$를 구한다.

$$D_i = \{\beta_1, \beta_2, ..., \beta_t\} \ \beta_{r_1} \le \beta_{r_2} \ , r_1 \le r_2 \qquad (4)$$
$$where \ \beta_k = \|x_{ij} - x_{kj}\|$$

식 (4)에서의 $x_{ij}$는 불완전 특징벡터 $x_i$의 데이터 값 중 분실 되지 않은 데이터 값을 의미하고, $x_{kj}$는 완전한 특징 벡터의 데이터 값이다.

단계 2 : 1 단계에서 계산된 거리를 이용하여 평균값에 사용할 $i$번째 불완전 벡터의 닮음 벡터 수 $u_i$를 결정한다.

$$u_i = |D_i|^* \gamma \tag{5}$$

본 논문에서는 식 (5) $\gamma$를 0.25로 설정 하였다.

단계 3 : 결정된 $u_i$의 수 만큼 대응되는 완전한 벡터의 값을 평균하여 불완전 벡터의 데이터 값으로 대체시킨다.

$$x_{i,missing} = \frac{1}{u_i}\sum_{k=1}^{u_i} x_{k,missing} \tag{6}$$

식 (6)의 $x_k$은 $\beta_k$에 대응되는 완전한 벡터를 의미한다.

단계 4 : (정의 2)의 $\alpha_{kj}$을 계산한다.

### 2.5 퍼지 분류기

새롭게 대체된 완전한 데이터 세트를 이용해 퍼지 분류기를 설계하여 실험하였다. 클래스 간 특징의 평균의 이용하여 삼각 퍼지 함수를 설계한다. 특징 벡터의 특징이 m가지 일 경우 m개의 삼각 퍼지 함수가 생성된다.

$$M_{ci} = \frac{1}{\eta_{ci}}\sum_{j=1}^{\eta_{ci}} x_{cij} \ where \ x_{cij} \in X_U \tag{7}$$

식 (7)을 이용하여 각 클래스의 특징별 평균을 구한다. $M_{ci}$ 는 $c$번째 클래스의 $i$번째 특징들의 평균이며, $\eta_{ci}$는 $c$번째 클래스의 $i$번째 특징의 데이터 수이다. 이때 불완전 데이터가 있을 경우 계산에 반영하지 않는다.

$$\begin{array}{ll} \mu_{ci}(x_j) = 0.1(x_j - M_{ci}) + 1 & \text{if } x_j < M_{ci} \\ \mu_{ci}(x_j) = -0.1(x_j - M_{ci}) + 1 & \text{if } x_j \geq M_{ci} \\ \mu_{ci}(x_j) = 0 & \text{if } \mu_{ci}(x) < 0 \end{array} \tag{8}$$

식 (8)과 같이 퍼지 함수를 설계한다. $\mu_{ci}$는 $c$번째 클래스의 $i$번째 특징의 소속도(membership) 함수이다.

테스트 셋에서의 찾아진 특징별 소속도 함수값과 effect factor $\alpha_{kj}$를 이용하여 식 (9)와 같이 가중치 평균으로 클래스를 결정한다. $\alpha_{kj}$를 사용함으로써 복원에서 올 수 있는 분류 에러율을 줄일 수 있다.

$$p_c = \frac{1}{\sum_{j=1}^{s}\alpha_{kj}}\sum_{j=1}^{s}\alpha_{kj}\mu_{cj}(x_j) \tag{9}$$

식 (9)의 $p_c$는 $x_j$벡터의 $c$번째 클래스의 소속도이며, 각 클래스의 소속도가 가장 큰 클래스로 분류한다.

## III. 실험 및 분석

실험을 위하여 UCI(University of Califonia, Irivne) machine learning repository 사이트[7]에서 제공되는 표준 데이터들 중 Breast Cancer Wisconsin 데이터와 Ecoli 데이터 세트를 이용하여 실험하였다. 각 데이터별 특징은 표 1과 같다.

표 1. 실험에 사용된 데이터별 특징

| | BCW | Ecoli |
|---|---|---|
| 클래스 수 | 2 | 5 |
| 특징 수 | 9 | 7 |
| 데이터 수 | 699 | 336 |
| 데이터 형식 | Integer | Real |
| Missing data | Yes(16) | No |

실험은 10-fold cross validation[8] 방식을 사용하였고, 데이터의 불완전 정도와 불완전 데이터를 복원하는 방식에 따라 세 가지 실험을 동시에 하여 비교하였다. 불완전 정도 $\delta$는 식 (10)을 이용하여 계산하였다.

$$\delta = \frac{n_u}{n_s} = \frac{|X_U|}{|X_A|^* s} \tag{10}$$

$\delta$는 불완전 정도에 따른 실험의 차이를 확인하기 위해 1.0, 0.95, 0.85, 0.75, 0.7에 대해서 UCI 데이터 별로 실험한다. $\delta$가 1.0일 경우 사실상 분실 데이터에 대한 복원 작업이 이루어지지 않음으로 모든 경우에 같은 결과를 확인할 수 있었다.

새로운 위치에서 랜덤하게 불완전 데이터를 생성하여 실험하기 때문에 10-fold cross validation 테스트 매회 마다 인식률이 달라진다. 따라서 실험은 $\delta$에 따라 결정되는 분실 데이터를 데이터 복원 기법에 따라 10회씩 10-fold cross validation 하여 그 평균값을 평균 인식률로 결정하였다. 각 실험에 대한 표기를 표 2와 같이 정의하였다.

표 2. 실험에 대한 표기

| 표기 | 실험 기법 |
|---|---|
| A | 기존의 불완전 데이터 처리 기법[6] |
| B | 본 논문에서 제안한 첫 번째 기법 |
| C | 본 논문에서 제안한 두 번째 기법 |

표 3. Breast Cancer Wisconsin 데이터를 이용한 실험

| $\delta$ | 불완전 데이터 수 | 인식률 | | |
|---|---|---|---|---|
| | | A | B | C |
| 1.0 | 0 | 95% | 95% | 95% |
| 0.95 | 307 | 93% | 95% | 94% |
| 0.85 | 922 | 93% | 96% | 93% |
| 0.75 | 1537 | 92% | 97% | 92% |
| 0.7 | 1844 | 92% | 95% | 90% |

표 3은 Breast Cancer Wisconsin 데이터를 이용한 실험이다. 모든 테스트에서 본 논문에서 제안하는 첫 번째 방법(B)이 가장 높은 인식률을 보이는 것을 확인 할 수 있다. 특히 B의 실험 결과에서 완전한 데이터보다도 더 높은 인식률을 확인할 수 있었다. 이러한 결과는 불완전한 데이터에서 완전한 데이터에 포함 된 분류에 방해가 되는 원 데이터들이 손실되고, 다시 가장 가까운 클래스의 평균값으로 대체 하는 것과 분류기 통과 시 $\alpha_{kj}$를 이용하여 가중치 평균함으로써 데이터 대체에서 생길 수 있는 에러율을 감소시키는 결과가 나타난 것으로 생각할 수 있다.

표 4. Ecoli 데이터를 이용한 실험

| $\delta$ | 불완전 데이터 수 | 인식률 | | |
|---|---|---|---|---|
| | | A | B | C |
| 1.0 | 0 | 86% | 86% | 86% |
| 0.95 | 114 | 83% | 85% | 84% |
| 0.85 | 343 | 77% | 82% | 79% |
| 0.75 | 572 | 71% | 78% | 76% |
| 0.7 | 687 | 69% | 76% | 71% |

표 4는 Ecoli 데이터를 이용한 실험 결과이다. 본 실험에서는 $\delta$이 작아짐에 따라서 인식률도 같이 작아짐을 알 수 있다. 하지만 앞의 실험과 마찬가지로 B 실험에서 가장 높은 인식률을 확인할 수 있었다.

## Ⅳ. 결 론

본 논문에서는 실생활에서 발생할 수 있는 불완전한 데이터를 이용하여 패턴 분류 하는 방법을 제안하였다. 기존의 Fuzzy C-Means에 사용되었던 클러스터링 방법을 패턴 분류에 적합한 방법으로 적용 하였고, 불완전 데이터를 복원하는 새로운 방법을 제시하여 기존의 방법 보다 높은 인식률을 확인 하였다. 본 논문은 완전한 데이터의 인식률을 높이는 것 보다는 불완전 데이터의 인식률을 얼마나 높일 수 있는 가에 초점을 두고 연구하였다. 그리하여 실제로 퍼지 분류기 또한 가장 기본적인 삼각 퍼지 함수를 사용하였다. 실제로 이와 같은 연구에서 완전한 데이터에 근접한 인식률을 얻을 수 있었다.

향후 연구 과제로는 Ecoli 데이터 실험에서의 인식률 저하 문제를 보완할 것이며 완전한 데이터에 대해서도 초점을 두어 클래스 내 통계 정보를 활용함으로써 더욱 합리적으로 클래스 소속도를 할당하는 기법을 보완하여 모든 데이터에 대해서 강인한 분류기를 설계하는 것과 다양한 표준 실험 데이터 세트를 이용하여 제안한 기법의 정당성과 일반성을 검증하는 것이 필요할 것으로 생각한다.

## 참고문헌

[1] K. B. Korb, A. E. Nicholson, Bayesian Artificial Intelligence, Chapman & Hall, 2004.
[2] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1955.
[3] V. Vapnik, Statical Learning Theory, John Wiley & Sons Inc., 1998.
[4] V.N. Vapnik, “An Overview of Statical Learning Theory,” IEEE Transactions of Neural Networks, Vol.10, No.5, pp.988-999, 1999.
[5] J. R. Quinlan, “C4.5:Program for Machine Learning,” Morgan Kaufmann, 1993.
[6] Zhiping Jia and Zhiqiang Yu “Fuzzy C-Means Clustering Algorithm Based on Incomplete Data,” IEEE International Conference on Information Acquisition, pp. 20-23, August 2006.
[7] A. Asunion and D. Newman, UCI machine learning repository, http://archive.ics.uci.edu/ml, School of Information and Computer Science, University of California, Irvine 2007.
[8] Ron Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp.1137-1143, 1995.