Fuzzy Classification Method for Processing Incomplete Dataset

Young Woon Woo, Kwang Eui Lee and Soowhan Han, *Member, KIMICS*

***Abstract* — Pattern classification is one of the most important topics for machine learning research fields. However incomplete data appear frequently in real world problems and also show low learning rate in classification models. There have been many researches for handling such incomplete data, but most of the researches are focusing on training stages. In this paper, we proposed two classification methods for incomplete data using triangular shaped fuzzy membership functions. In the proposed methods, missing data in incomplete feature vectors are inferred, learned and applied to the proposed classifier using triangular shaped fuzzy membership functions. In the experiment, we verified that the proposed methods show higher classification rate than a conventional method.**

***Index Terms* — Fuzzy Classifier, Incomplete Dataset, Triangular Fuzzy membership function, Weight**

I. INTRODUCTION

Pattern recognition is defined as "a research field of artificial intelligence handling problems for recognizing objects by a device having computational capability". Various techniques have been developed for classifying and processing abundant data in real world problems. In the various techniques, the recognition technique is one of the most typical techniques for searching and classifying information.

It is important and essential to classify huge amount of information effectively in information processing fields. But in the real world problems, some data may be lost and distorted according to situations, and these data can lower the performance of a classifier and also make the problem more difficult.

In order to manage such incomplete data, many techniques such as Bayesian classifier [1], Support Vector Machine (SVM) [2-4], and a method by J. Ross Quinlan[5] and a clustering method by fuzzy c-means[6], were proposed.

Manuscript received July 1, 2010; accepted July 1, 2010.

Young Woon Woo is with the Dept. of Multimedia Eng., Dong-Eui University, Busan, Korea (e-mail: ywwoo@deu.ac.kr). (Corresponding Author)
Kwang Eui Lee is with the Dept. of Multimedia Eng., Dong-Eui University, Busan, Korea (e-mail: kelee@deu.ac.kr).
Soowhan Han is with the Dept. of Multimedia Eng., Dong-Eui University, Busan, Korea (e-mail: swhan@deu.ac.kr).

In this paper, we proposed two new methods by improvement of unsupervised fuzzy c-means clustering method using supervised learning of fuzzy classifier, and applied the proposed methods to a pattern classification problem. In the experiment, Ecoli dataset from UCI machine learning repository site [7], was used for the proposed methods and the results are compared and analyzed.

II. THE PROPOSED FUZZY CLASSIFICATION METHODS

A. Replacement of Missing Data

In this paper, we estimate and recover missing feature values of incomplete feature vectors, and then apply the recovered feature vectors to fuzzy classifier. In order to recover missing feature values, we used algebraic average of feature values in each class. The instance of incomplete feature vector can be expressed as (?, 2, 3, 4) – 25% missing feature vector and (?, 2, ?, 4) – 50% missing feature vector when the complete feature vector is (1, 2, 3, 4).

In this paper, we used some definitions [6] for describing the conventional classification method and the proposed classification methods.

- *Definition #1:*
  Incomplete data X_A

$$X_A = \{x_1, x_2, ..., x_n\} \subset R^n$$
$$x_k = \{x_{k1}, x_{k2}, ..., x_{ks}\} \subset R^s$$

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

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

- *Definition #2:*
  The effect factor α_{kj}

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

- *Definition #3:*
  The resembling factor β_{ij}

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

B. The Conventional Classification Method

In the conventional clustering method [6], missing feature values are replaced by other complete feature values from the most similar feature vector using Euclidean distance between the incomplete feature vector and other all of complete feature vectors.

In this paper, we modified the conventional method for solving pattern classification problems and proposed two kinds of methods for classifying incomplete feature vectors.

C. The First Proposed Classification Method

The first proposed classification method is as follows.

Step 1: Calculate algebraic centers of each class using Equation (1).

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

In equation (1), c_k means center of each class, the centers of each class are calculated except missing feature values. η_{ki} is the number of complete feature values in i^{th} feature vector of k^{th} class.

Step 2: β_k', the distance between arbitrary c_k and missing feature vector is calculated by equation (2).

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

cn means the number of classes, x_j means complete feature values in incomplete feature vector x.

Step 3: Missing feature values are replaced by the center value of minimum distance by equation (3).

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

Step 4: Calculate α_{kj} of *Definition #2*.

D. The Second Proposed Classification Method

The second proposed classification method is to replace missing data by mean values of similar data from all of data in each class. The description of the second proposed classification method as follows.

Step 1: Calculate distance set D_i between i^{th} incomplete feature vector and all of feature vectors using equation (4).

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

In equation (4), x_{ij} means complete feature values in incomplete feature vector x_i and x_{kj} means complete feature vector.

Step 2: Calculate similarity degree of i^{th} incomplete vector using distance set from Step 1 and equation (5).

$$u_i = |D_i| * \gamma \qquad (5)$$

In this paper, we used 0.25 as γ value.

Step 3: Missing feature values are replaced by mean values of complete vectors corresponding to determined number of u_i.

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

In equation (6), x_k means complete vector corresponding to β_k.

Step 4: Calculate α_{kj} of *Definition #2*.

E. Fuzzy Classifier Used in the Proposed Methods

We designed a fuzzy classifier using newly acquired complete data by replacement. Fuzzy membership functions are made by mean values of each class. When the number of feature values of each vector is m, m fuzzy membership functions are defined.

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

We calculated mean values of each class using equation (7). In equation (7), M_{ci} means mean value of i^{th} feature value of c^{th} class, η_{ci} means the number of i^{th} feature vector of c^{th} class. We excluded incomplete data in equation (7).

We made fuzzy membership functions using equation (8). In equation (8), μ_{ci} means membership value of i^{th} feature value of c^{th} class.

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

The final classification results are decided by equation (9).

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

In equation (9), p_c means membership value of x_j vector of c^{th} class and the largest membership value from each membership grades of each class is the resultant classification result.

III. EXPERIMENT AND ANALYSIS

We carried out the experiment on the two proposed methods by 'Ecoli' standard dataset and Breast Cancer Wisconsin(BCW) from UCI(University of California Irvine) machine learning repository [7]. The characteristics of the 'Ecoli' dataset and 'BCW' dataset are shown in Table I.

TABLE I
CHARACTERISTICS OF 'ECOLI' DATASET USED IN THE EXPERIMENT

	Ecoli	BCW
Number of Classes	5	2
Number of Features	7	9
Number of samples	336	699
Data Format	Real	Integer
Missing Data	No	Yes(16)

We compared the proposed methods by different degrees of completeness. The degree of completeness (δ) is calculated by equation (4).

$$\delta = \frac{n_u}{n_s} = \frac{|X_U|}{|X_A| * s} \tag{4}$$

We used 5 kinds of δ such as 1.0, 0.95, 0.85, 0.75, 0.7. δ of 1.0 means all of feature values are complete, no missing. δ of 0.7 means 70 percent of feature values are complete. The performance of the proposed method is measure by 10-fold cross validation method [8]. Table II shows the notation for each method for experiment and Table III and Table IV show the experimental results of the conventional method and the proposed methods by different δ values.

TABLE II
NOTATION FOR EACH METHOD

Notation	Experimented Method
A	The Conventional Method [6]
B	The First Proposed Method
C	The Second Proposed Method

TABLE III
EXPERIMENTAL RESULTS FOR ECOLI DATASET

δ	The Number of Incomplete Feature Values	Classification Rate		
		A	B	C
1.0	0	86%	86%	86%
0.95	307	83%	85%	84%
0.85	922	77%	82%	79%
0.75	1537	71%	78%	76%
0.7	1844	69%	76%	71%

TABLE IV
EXPERIMENTAL RESULTS FOR BCW DATASET

δ	The Number of Incomplete Feature Values	Classification Rate		
		A	B	C
1.0	0	95%	95%	95%
0.95	114	93%	95%	94%
0.85	343	93%	96%	93%
0.75	572	92%	97%	92%
0.7	687	92%	95%	90%

In Table III and Table IV, we could verify the classification rate decreased in proportional to diminishing δ values.

IV. CONCLUSIONS

In this paper, we proposed two methods to classify incomplete dataset appeared in real world problems. In the proposed methods, a conventional clustering method by fuzzy c-means is modified and applied to pattern classification method and two new methods for replacing missing feature values with appropriate values, is presented. In the experiment using standard pattern classification dataset from UCI machine repository site, we verified the proposed methods are more efficient than the conventional method in classification rate.

REFERENCES

[1] K. B. Korb and 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.

Young Woon Woo
Received the B.S. degree, M.S. degree and Ph.D. degree in electronic engineering from Yonsei University, Seoul, Korea in 1989, 1991 and 1997, respectively. Since 1997, he has been a professor in Department of Multimedia Eng., Dong-Eui University, Busan, Korea. His research interests are in the area of artificial intelligence, image processing, pattern recognition and medical information.

Kwang Eui Lee
Received his B.S., M.S. and the Ph.D. degrees from Sogang University, Seoul, Korea in 1990, 1992, and 1997, respectively. From 1997 to 2001, he joined ETRI as a senior research member. Since 2001, He has been an associate professor of Dongeui University. His research interests include computation theory, artificial life, context awareness and their applications

Soowhan Han
Member of KIMICS. Received B.S. degree in electronics, Yonsei University, Korea, in 1986, and M.S. and Ph.D. degree in Electrical & Computer Eng., Florida Institute of Technology, U.S.A. in 1990 and 1993, respectively. From 1994 to 1996, he was an assistant professor of the Dept. of Computer Eng., Kwandong University, Korea. In 1997, he joined the Dept. of Multimedia Eng., Dongeui University, Korea, where he is currently a professor. His major interests of research include Digital Signal & Image Processing, Pattern Recognition and Neural Networks.