Fuzzy Clustering-Based on Aggregate Attribute Method

Jia-Wen Wang and Ching-Hsue Cheng

Department of Information Management,
National Yunlin University of Science and Technology,
123, section 3, University Road, Touliu, Yunlin 640, Taiwan, R.O.C.
{g9220803, chcheng}@yuntech.edu.tw

Abstract. This paper, we propose a fuzzy clustering-based on aggregate attribute method for classification tasks, which comprises three phases: (1) Calculate the aggregate attribute values. (2) Apply fuzzy clustering to cluster the aggregate values. (3) Predict the testing data's class. For verifying proposed method, we use two datasets to illustrate our performance, the datasets are: (1) Iris; (2) Wisconsin-breast-cancer dataset. Finally, we compare with other methods; it is shown that our proposed method is better than other methods.

1 Introduction

To assign new instances from a domain to one class of mutually exclusive classes based on the observed attributes of the instance is a common problem that occurs in the sciences, social sciences and business. The increasing complexity and dimensionality of classification problems, it is necessary to deal with structural issues of the identification of classifier systems. Selecting the important attributes and determining effective initial discretization of the input domain are important tasks [2, 7]. But the problems are difficulty and complexity problem in the real world.

In this paper, we propose a fuzzy clustering-based on aggregate attribute method for classification tasks. An aggregate attribute value is composed of a set of attributes [15]. We obtain aggregate value by beta coefficient, and use fuzzy c-means to build clusters by aggregate values. In order to verify our method, we use two databases: (1) Iris dataset; (2) Wisconsin-breast-cancer dataset. Moreover, estimating accuracy is important that it allows one to evaluate how accurately a given classifier will label future data, that is, data on which the classifier has not been trained [7]. Therefore, we partition the dataset as training/testing data for estimating accuracy. Finally, the result presented the accuracy of the proposed method is better than the existing methods.

The rest of this paper is organized as follows. In Section 2, we briefly review literature. In section 3, we describe the proposed method in detail. In section 4, we present some actual example to verify our method and compare with other methods. Finally, section 5 is conclusion.

2 Preliminary

In this section, we describe about the fuzzy clustering method and regression, and beta coefficient.

M. Ali and R. Dapoigny (Eds.): IEA/AIE 2006, LNAI 4031, pp. 478 – 487, 2006.

2.1 Fuzzy Cluster Method

Clustering has been obtaining popularity as an efficient tool of data analysis to understand and visualize data structures. The different types of clustering algorithms can be classified into hierarchical, partitional, categorical and large DB. In partitional clustering algorithms, there are many clustering algorithm such as Squared error clustering algorithm, K-means clustering, PAM (partitioning around medoides) algorithm, Bond energy algorithm (BEA), clustering with genetic algorithms, clustering with neural networks etc [7].

The prevalent formulation of this task is to use C feature vectors $v_j (j = 1, 2, ..., C) \in R^C$ to represent the C clusters such that a sample x_k is classified into the jth cluster according to some measure of similarity and its corresponding objective function. Two types of clustering techniques are usually of consideration, namely, hard (or crisp) clustering and fuzzy (or soft) clustering. Formally, hard clustering divides U into subsets $A_1, A_2, ..., A_C$, such that $A_1 \cup A_2 \cup ... \cup A_C = U$ and $A_i \cap A_j = \varnothing, \forall i \neq j, i, j = 1, 2, ..., C$. On the contrary, fuzzy clustering derives a number of subsets $A_1, A_2, ..., A_C$ of U such that $A_1 \cup A_2 \cup ... \cup A_C = U$ and $A_i \not\subset A_j, \forall i \neq j, i, j = 1, 2, ..., C$.

Fuzzy C Mean (FCM), is the most famous and basic fuzzy clustering algorithm. FCM attempts to find a fuzzy partition of the data set by minimizing the following within group least-squares error objective function with respect to fuzzy memberships u_{ik} and center v_i:

$$J_m(X, U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m d^2(x_k; v_i) \tag{1}$$

where $m > 1$ is the fuzziness index used to tune out the noise in the data, n is the number of feature vectors x_k, $c > 2$ is the number of clusters in the set and $d(x_k; v_i)$ is the similarity measure between a datum and a center. Minimizing J_m under the following constraints:

$$\begin{aligned} &(1) 0 \le u_{ik} \le 1, \forall i, k, \\ &(2) 0 < \sum_{k=1}^{n} u_{ik} \le n, \forall i, \\ &(3) \sum_{i=1}^{c} u_{ik} = 1, \forall k, \end{aligned} \tag{2}$$

yields an iterative minimization pseudo-algorithm well known as the FCM algorithm. The components v_{il} of each center v_i and the membership degrees u_{ik} are updated according to the expressions

$$v_{il} = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_{kl}}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{3}$$

and

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d(x_k; v_i)}{d(x_k; v_j)}\right)^{2/m-1}} \tag{4}$$

The membership matrix $U(c,n)$ is initialized randomly or by defining $U^{(0)}(c,n)$ as follows:

$$U^{(0)} = (1 - \frac{\sqrt{2}}{2})U_u + \frac{\sqrt{2}}{2}U_r \tag{5}$$

where $U_u = [1/c]$ and U_r is a random hard partition of data.

2.2 Regression

In multiple regression analysis, we study the relationships among three or more variables. We write the multiple regression model as follows [6]:

$$y_i = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + e_j \tag{6}$$

where y_i is a typical value of Y, the dependent variable, from the population of interest; $\beta_0, \beta_1, \ldots, \beta_k$ are the population partial regression coefficients; and $x_{1j}, x_{2j}, \ldots, x_{kj}$ are observed values of the independent variables $X_1, X_2, \ldots, X_k$, respectively. e_j are assumed to be random and normal distribution $N(0, \sigma^2)$, $j = 1,2,\cdots,n$.

2.2.1 Beta Coefficient

The standardized the coefficients are called beta coefficients. The steps of the beta coefficients are [6]:

(1) Standardized all of your variables $y_1 = (y - \bar{y})/S_y$, denote y_1 is the standardized of the y, $\bar{y}$ is the mean of the y. and S_y is the standard deviation of y.

(2) Assume z_1 is standardized of the X_1, z_2 is standardized of the X_2.

(3) Then, we can obtain the regression model. $y_1 = \beta_1 \times z_1 + \beta_2 \times z_2$.

The advantage of Beta coefficients (as compared to regression coefficients which are not standardized) is that the magnitude of these Beta coefficients allows you to compare the relative contribution of each independent variable in the prediction of the dependent variable.

3 Fuzzy Clustering-Based on Aggregate Attribute Method

In this section, we proposed a method based on fuzzy clustering technique and the aggregate attribute. Assume there are n attributes var_1, $var_2, ..., var_n$. We use the beta coefficient to calculate the aggregate values, and then we use the fuzzy clustering method to cluster the value. The concept is shown in figure 1.The advantage of Beta coefficients is: (1) eliminates the problem of dealing with different units of measurement (2) allows you to compare the relative contribution of each independent variable in the prediction of the dependent variable. The section 3.1 is the steps of this process, and the section 3.2 is the algorithm for the proposed method.

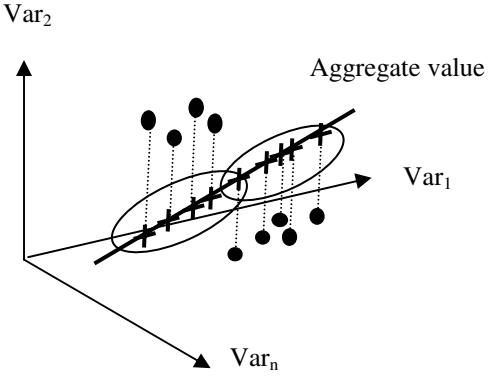

Fig. 1. The concept of the proposed method

The steps of this process are described below:

A. ***Data selection & transformation:*** Extract features (Selection input variable and class variable), transform or consolidate into appropriate forms. The fundamental task of variable selection must be performed by the analyst. Moreover, the selection of a dependent variable is many times dictated by the research problem and the research target.

B. ***Calculate the aggregate value:*** Different input variables have different influence on the class variable. We calculate the aggregate value between the attributes by Beta coefficient.

C. ***Cluster the aggregate value:*** Use the fuzzy cluster method to cluster the aggregate value.

D. ***Evaluation:*** Compare accuracy with the past research.

3.1 The Algorithm for Proposed Method

In this section, we present the algorithm for proposed method. Assume an n tuples relation database including 6 attributes is shown as Table 1. The attributes "X_1", "X_2", "X_3" and "X_4" are called input variables and the attribute "Class" is called a class variable. And the NO is the tuple's number.

Table 1. A relation database

NO	X_1	X_2	X_3	X_4	Class
O_1	X_{11}	X_{21}	X_{31}	X_{41}	C_1
O_2	X_{12}	X_{22}	X_{32}	X_{42}	C_2
⋮	⋮	⋮	⋮	⋮	⋮
O_n	X_{1n}	X_{2n}	X_{3n}	X_{4n}	C_n

Step 1: *Data selection & transformation:*
We use the 50% dataset as training data, and the 50% as testing data.

Step 2: *Calculate the aggregate value by training data:*
In this paper, we use the beta coefficient to calculate the aggregate attribute value.

$$\text{The aggregate value}_{\text{n}} = \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_i x_{in} \tag{7}$$

Where the β_i is beta coefficient of the *i*th input variable, X_{in} is the *i*th input variable. n is the number of tuple in database.

Step 3: *Building clusters:*
Use fuzzy c-means to build clusters by aggregate values. This step is to build clusters by aggregate values. Assume that we obtain k clusters (i.e., $C_1, C_2, \ldots, C_k$) shown as below:

$$C_1 = \{a_{1,1}, a_{1,2}, \ldots, a_{1,P_1}\} \cdots C_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,j}\} \cdots C_k = \{a_{k,1}, a_{k,2}, \ldots, a_{k,j}\} \tag{8}$$

where a_{ij} denotes the jth aggregate value of C_i.

Step 4: *Predict the testing data's class:*
Calculate the Euclidean distance $Dist_i$ between the tuple of testing data and C_i.

$$Dist_i = \sqrt{\left(C_{i\_center} - a_{ij}\right)^2} \tag{9}$$

where C_{i_center} is the ith cluster center, and the minimal $Dist_i$ denotes "The tuple belongs to C_i."

Step 5: *Evaluation:* Compare accuracy with the past research.
For verifying proposed method, we use two public datasets to illustrate our performance, the datasets are: (1) Iris dataset; (2) Wisconsin-breast-cancer dataset.

4 Numerical Simulations and Results

For verification and comparison, we use two datasets: (1) Iris dataset; (2) Wisconsin-breast-cancer dataset. The two examples are the well-known classification problem, and the data is freely available in the internet [4, 10, 11].

4.1 Iris Dataset

There are three species (or class labels) in the class variable: setosa, versicolor, and virginica. From each species there are 50 observations for sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW) in cm. The dataset is shown in Table 2. (1)From step 1 we partition the dataset, the 50% dataset as training data and others as testing data. (2) Secondly, we obtain aggregate value (see Table 4) from the proposed method. The beta coefficient (see Table 3) is shown in table. (3) Thirdly, use fuzzy c-means to build clusters by aggregate values. We can obtain the cluster center. It is shown in Table 5. (4) Finally, we can use the training's cluster to predict the testing data's class by equation (9).

The experimental result is shown in Table 6. The proposed model is compared with Chen and Yu's algorithm [5], Wu and Chen's algorithm [16], and Hong and Lee's algorithm [8]. From Table 6, the results indicate the proposed algorithm is the best performance in testing data, and the accuracy rate is 97.3333%.

Table 2. Iris dataset

Data no	Sepal length	Sepal width	Petal length	Petal width	Class
1	5.1	3.5	1.4	0.2	1
2	4.9	3	1.4	0.2	1
		⋮			
149	6.2	3.4	5.4	2.3	3
150	5.9	3	5.1	1.8	3

Note. The class values: 1 is setosa; 2 is versicolor; 3 is virginica.

Table 3. Beta coefficient

Variables	Beta coefficients
Sepal_length	-0.11092
Sepal_width	-0.02342
Petal_length	0.488904
Petal_width	0.568151

Note. [a.] Dependent Variable: Class.

Table 4. Aggregate attribute

Data no	sepal length	sepal width	petal length	petal Width	Aggregate value
1	5.1	3.5	1.4	0.2	-0.11121
2	4.9	3	1.4	0.2	-0.06234
...	...	...	...	...	...
74	6.3	2.7	4.9	1.8	2.332041
75	6.7	3.3	5.7	2.1	2.800344

Table 5. Cluster center

Fuzzy c-means cluster center
-0.03307
1.773493
2.661606

Table 6. Comparison results with other methods for iris dataset

Algorithms	Classification accuracy rate
The proposed algorithm (75 training instances and 75 testing instances)	**97.3333**%
Chen and Yu's algorithm (75 training instances and 75 testing instances)	96.3427%
Wu and Chen's algorithm (75 training instances and 75 testing instances)	96.2100%
Hong and Lee's algorithm (75 training instances and 75 testing instances)	95.5700%

4.2 Wisconsin-Breast-Cancer Dataset

The second numerical example concerns the determination of the breast cancer in humans from Wisconsin University hospital in Madison, USA. This is also freely available in the Internet via anonymous ftp from ics.uci.edu in directory/ pub/ machine-learning-databased/ Wisconsin-breast-cancer. This problem has 9 features/input attributes which determine whether a patient is benign or malign. The 9 features are: (a) Clump Thickness (b) Uniformity of Cell Size (c) Uniformity of Cell Shape (d) Marginal Adhesion (e) Single Epithelial Cell Size (f) Bare Nuclei (g) Bland Chromatin

(h) Normal Nucleoli (i) Mitoses. There are 683 samples/patterns. This data has been used in the past by Mangasarian et al. [10, 11] and Bennet et al. [4]. The dataset is shown in Table 7. The beta coefficient is shown in Table 8.

We use the 50% dataset as training data, and the 50% as testing data. The result of the breast cancer classification problem is shown in Table 9. As the error estimates are either obtained from 10-fold cross validation or from testing the solution once by using the 50% of the data as training set. The related methods [1, 12, 13, 14] are compared with the proposed model. From Table 9, the proposed model is the best performance and the accuracy rate is 98.2%.

Table 7. The Wisconsin-breast-cancer dataset

ID	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
63375	9	1	2	6	4	10	7	7	2	2
128059	1	1	1	1	2	5	5	1	1	1
					⋮					
160296	5	8	8	10	5	10	8	10	3	2

Note. The class values: 1 is benign, 2 is malign.

Table 8. The beta coefficient

Variables	Beta Coefficients
Clump Thickness	0.26101
Uniformity of Cell Size	0.032392
Uniformity of Cell Shape	0.143727
Marginal Adhesion	0.094532
Single Epithelial Cell Size	0.074322
Bare Nuclei	0.345085
Bland Chromatin	0.03087
Normal Nucleoli	0.09801
Mitoses	0.013683

Note. [a] Dependent Variable: Class.

Table 9. Comparison results with other methods for wisconsin-breast-cancer dataset

Algorithms	Classification accuracy rate
The proposed model (342 testing instances)	**98.2%**
Konstam's algorithm (342 testing instances)	97.5%
Setiono's algorithm (2000)	97.36
Setiono's algorithm (2000)	98.1
Pena-Reyes and Sipper(2000)	97.36
Pena-Reyes and Sipper(2000)	97.07
Nauck and Kruse(1999)	95.06

5 Conclusions

For solving classification problems, we have proposed a fuzzy clustering-based on aggregate attribute method. From two datasets' experiment results; we can see that the accuracy of the proposed method is better than the existing methods. That is, the proposed method can get better estimated accuracy than the existing methods. In the future, we could consider other aggregate attribute methods to improve the performance.

Acknowledgements

The authors would like to thank: (1) the University of Wisconsin Hospitals, Dr. William H. Wolberg, provided the dataset. (2) the anonymous referees whose comments have improved the presentation of the paper.

References

1. Konstam, A., Group classification using a mix of genetic programming and genetic algorithms, ACM Press, (1998)
2. Abonyi, J., Roubos, J.A., Szeifert, F. Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization, International Journal of Approximate Reasoning 32, (2003) 1-21
3. Arnold, S.F.: Mathematical Statistics, Prentice-Hall, Englewood Cliffs NJ (1990)
4. Bennett, K.P. & Mangasarian, O.L., Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software 1 (1992) 23±34.
5. Chen, S.M., & Yu, C.A., new method to generate fuzzy rules from training instances for handling classification problems. Cybernetics and Systems: An International Journal, 34(3), (2003) 217-232.

6. Draper, N.R., Smith, H.: Applied Regression Analysis, John Wiley and Sons, New York (1998).
7. Han, J. & Kamber, M.: Data Mining: Concepts and Techniques, Morgan Kaufmann (2001)
8. Hong, T. P., & Lee, C. Y. Induction of fuzzy rules and membership functions from training examples. Fuzzy Sets and Systems, 84(1), (1996) 33-47.
9. MacQueen, JB.: Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, (1967) 281-297.
10. Mangasarian, O.L. & Wolberg, W.H., Cancer diagnosis via linear programming, SIAM News 23 (1990) 1-18.
11. Mangasarian, O.L., Setiono, R. and Wolberg, W.H. Pattern recognition via linear programming: Theory and application to medical diagnosis, in: T.F. Coleman, Y. Li (Eds.), Large-scale numerical optimization, SIAM Publications, Philadelphia, 1990, pp. 22±30.
12. Nauck, D., Kruse, R. Obtaining interpretable fuzzy classification rules from medical data, Artif. Intell. Med. 16, (1999) 149-169.
13. Peña-Reyes, C.A., Sipper, M. A fuzzy genetic approach to breast cancer diagnosis, Artif. Intell. Med. 17, (2000) 131-155.
14. Setiono, R., Generating concise and accurate classification rules for breast cancer diagnosis, Artif. Intell. Med. 18, (2000) 205-219.
15. Subtil, P., Mouaddib, N. and Foucaut O. A fuzzy information retrieval and management system and its applications, Proceedings of the 1996 ACM symposium on Applied Computing, (1906) 537-541.
16. Wu, T. P., & Chen, S.M., A new method for constructing membership functions and fuzzy rules from training examples. IEEE Transactions on Systems, Man and Cybernetics-Part B, 29(1), (1999) 25-40.