A Preprocessed Induced Partition Matrix Based Collaborative Fuzzy Clustering For Data Analysis

M. Prasad
Dept. of Computer Science
National Chiao Tung University
Hsinchu, Taiwan
mukeshnctu.cs99g@nctu.edu.tw

D. L. Li
Dept. of Computer Science
National Chiao Tung University
Hsinchu, Taiwan
lazybones000@gmail.com

Y. T. Liu
Dept. of Electrical Engineering
National Chiao Tung University
Hsinchu, Taiwan
tingting76319@gmail.com

L. Siana
Dept. of Electrical Engineering
National Chiao Tung University
Hsinchu, Taiwan
linda.siana@gmail.com

C. T. Lin
Dept. of Electrical Engineering
National Chiao Tung University
Hsinchu, Taiwan
ctlin@mail.nctu.edu.tw

A. Saxena
Dept. of Computer Science & IT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India
amitsaxena65@rediffmail.com

Abstract—**Preprocessing is generally used for data analysis in the real world datasets that are noisy, incomplete and inconsistent. In this paper, preprocessing is used to refine the inconsistency of the prototype and partition matrices before getting involved in the collaboration process. To date, almost all organizations are trying to establish some collaboration with others in order to enhance the performance of their services. Due to privacy and security issues they cannot share their information and data with each other. Collaborative clustering helps this kind of collaborative process while maintaining the privacy and security of data and can still yield a satisfactory result. Preprocessing helps the collaborative process by using an induced partition matrix generated based on cluster prototypes. The induced partition matrix is calculated from local data by using the cluster prototypes obtained from other data sites. Each member of the collaborating team collects the data and generates information locally by using the fuzzy c-means (FCM) and shares the cluster prototypes to other members. The other members preprocess the centroids before collaboration and use this information to share globally through collaborative fuzzy clustering (CFC) with other data. This process helps system to learn and gather information from other data sets. It is found that preprocessing helps system to provide reliable and satisfactory result, which can be easily visualized through our simulation results in this paper.**

Keywords— fuzzy c-means (FCM); collaborative fuzzy clustering (CFC); preprocessing; privacy and the security.

I. Introduction

The Fuzzy c-means (FCM) method has been studied since 1981 [1, 2] and it has been also improved in different ways and applied in various research areas and many researchers gave contribution to it from time to time leading to different variants of FCM. Now fuzzy clustering has become a matured method in the field of unsupervised clustering. This method is suitable to handle one dataset at a time. The main objective of fuzzy clustering is to assure that it operates not only on data, but takes full advantage of various sources of knowledge which comes from different sources of available data when dealing with the problem at hand. To do the clustering with more than one dataset and find some similar properties among these data sets, we start looking for some kind of collaboration process, but because of privacy and security issues datasets are not allowed to do collaboration directly. Taking these issues into account, the concept of collaborative clustering [3] was introduced. In this clustering algorithm, several subsets of patterns can be processed together with an objective of finding a structure that is common to all of them. All datasets are clustered locally by FCM and go for collaboration globally.

Collaborative fuzzy c-means clustering was introduced by Pedrycz [3-4] and later on, this work was carried by proposing an induced partition matrix [5, 6]. However, there is a problem in previously proposed methods; in the objective functions, direct subtraction of partition matrices of two different datasets was computed without taking into consideration the properties of the datasets. According to the mathematical meaning of matrix subtraction, each row of each matrix should contain the same properties and this property is violated in previously proposed method. Without consideration of this property, the meaning of subtraction will be changed and the result will give different direction to the system and it may cause, misguiding and poor performance of systems. So by doing preprocessing before collaboration, will help system to provide better result and to know better about different data sets and learn its behavior.

The rest of the paper is organized as follows: Section II gives a simple introduction of FCM, collaborative fuzzy clustering and introduce the preprocessing technique with refined collaboration. Section III shows the experimental results on Iris data and Mackey glass time series data and compare results without preprocessing and with preprocessing and finally the conclusions are covered in Section IV.

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

II. FCM, CFC and Proposed Algorithm

A. Fuzzy C-Means Clustering (FCM)

In 1981, Bezdek introduced a data clustering technique called fuzzy c-means (FCM), which allows each data point belong to one or more clusters that is specified by a membership function. The minimization of objective which decide the performance of FCM is defined as.

$$J_M=\sum_{i=1}^{N}\sum_{j=1}^{c}u_{ij}^{m}\| x_i-v_j \|^2 \tag{1}$$

where, M is any real number great than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i-th of d-dimension data, v_i is the d-dimension of the cluster and $\|*\|$ is any norm expressing the similarity between any measured data and the center.

B. Procedure for FCM

1. Set up a value of c (number of cluster);
2. Select initial cluster prototype $V_1,V_2,\ldots\ldots,V_c$ from X_i , $i=1,2,\ldots\ldots,N$;
3. Computer the distance $\|X_i-V_j\|$ between objects and prototypes;
4. Computer the elements of the fuzzy partition matrix ($i=1,2,\ldots\ldots,N$; $j=1,2,\ldots\ldots,c$)

$$u_{ij}=\left[\sum_{l=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_l\|}\right)\right]^{-1} \tag{2}$$

5. Compute the cluster prototypes ($j=1,2,\ldots\ldots,c$)

$$V_j=\frac{\sum_{i=1}^{N}u_{ij}^{2}x_i}{\sum_{i=1}^{N}u_{ij}^{2}} \tag{3}$$

6. Stop if the convergence is attained or the number of iterations exceeds a given limit. Otherwise, go to step 3

C. Collaborative Fuzzy Clustering (CFC)

Collaborative fuzzy c-means clustering was introduced by Pedrycz [4]. Basically collaborative clustering has its two typical forms called horizontal collaborative clustering and vertical collaborative clustering. In this paper, horizontal collaborative clustering has been worked with. In our previous work [7, 8], we applied vertical collaborative clustering for EEG Data and horizontal collaborative for designing and modeling a system with Mamdani type fuzzy inference system. The general scheme of horizontal collaborative clustering and vertical collaborative clustering are shown in Fig. 1 and Fig. 2 respectively.

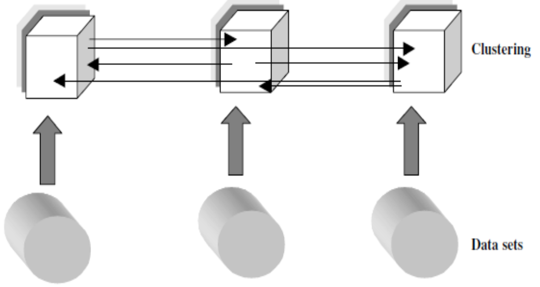

Fig. 1. A General scheme of horizontal clustering

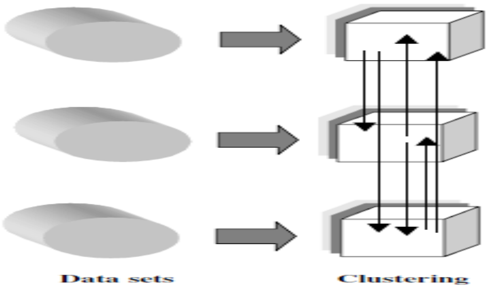

Fig. 2. A General scheme of vertical clustering

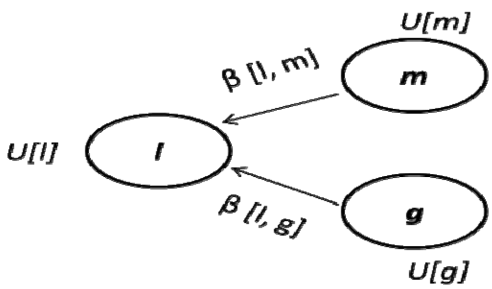

Fig. 3. Collaborative clustering scheme

The objective function for collaboration technique is explained as:

$$\begin{aligned}Q[l]&=\sum_{i=1}^{N}\sum_{j=1}^{c}u_{ij}^{2}[l]d_{ij}^{2}[l]\\&+\sum_{\substack{m=1\\m\neq l}}^{p}\beta[l,m]\sum_{i=1}^{N}\sum_{j=1}^{n}\{u_{ij}[l]-u_{ij}[m]\}^{2}d_{ij}^{2}[l]\end{aligned} \tag{4}$$

where, β is a user defined parameter based on datasets (β >0), $\beta[l,m]$ denotes the collaborative coefficient with collaborative effect on dataset l through m, c is a number of cluster. $l=1,2,\ldots\ldots,P$. P is a number of datasets, N is the number of patterns in the dataset, u represents the partition matrix, n is a number of features, and d is an Euclidean distance between patterns and prototypes.

Fig. 3 shows the connections of matrices in order to accomplish the collaboration between the subsets of the database. The optimization of $Q[l]$ as shown in (4) and (6) involves the determination of the partition matrix $u[l]$ and the prototypes $v_i[l]$. Prototype and partition matrices bring the way of structural findings at the each dataset. The reason why induced partition matrices are introduced is because of the focus on the partition matrices as one of its components to be adjusted to FCM optimization. The induced partition matrices are calculated based on (5) with local data and cluster prototype sent from the other data sites.

$$\tilde{u}_{ij} = \left[\sum_{l=1}^{c} \left(\frac{\| X_i - V_j \|}{\| X_i - V_l \|} \right)^2 \right]^{-1} \tag{5}$$

The optimization equation (6) is introduced by Pedrycz [5] in 2008 by introducing a new term called induced partition matrix. First we solve the problem for each data set separately and allow the results to interact globally by forming a collaborative process between the data sets. Apply Lagrange multipliers to minimize the objective function with respect to the partition matrix due to the standard constraints imposed on the partition matrix. Collaborative fuzzy partitioning is carried out through an iterative optimization of the objective function as shown above in (6) with an update of partition matrix $u[l]$ and the prototype $v_i[l]$. For optimization details please refer [3, 4].

$$\begin{aligned} Q[l] &= \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^2[l] d_{ij}^2[l] \\ &+ \sum_{\substack{m=1 \\ m \neq l}}^{p} \beta[l,m] \sum_{i=1}^{N} \sum_{j=1}^{n} \{u_{ij}[l] - \tilde{u}_{ij}[l/m]\}^2 d_{ij}^2[l] \end{aligned} \tag{6}$$

Minimize (6) at each data site by iteratively proceeding with the iterative calculations of the partition matrix and the prototypes, that is

$$u_{st}[l] = \frac{\varphi_{st}[l]}{1+\psi[l]} + \frac{1}{\sum_{j=1}^{c} \frac{d_{st}^2[l]}{d_{jt}^2[l]}} \left[1 - \sum_{j=1}^{c} \frac{\varphi_{jt}[l]}{1+\psi[l]} \right] \tag{7}$$

where,

$$\varphi_{st}[l] = \sum_{\substack{m=1 \\ m \neq l}}^{p} \beta[l,m] u_{st}[m] \tag{8}$$

$$\psi[l] = \sum_{\substack{m=1 \\ m \neq l}}^{p} \beta[l,m] \tag{9}$$

and

$$v_{st}[l] = \frac{\sum_{k=1}^{N} u_{sk}^2[l] x_{kt}[l] + \sum_{\substack{m=1 \\ m \neq l}}^{p} \beta[l,m] \sum_{k=1}^{N} (u_{sk}[l] - u_{sk}[m])^2 x_{kt}[l]}{\sum_{k=1}^{N} u_{sk}^2[l] + \sum_{\substack{m=1 \\ m \neq l}}^{p} \beta[l,m] \sum_{k=1}^{N} (u_{sk}[l] - u_{sk}[m])^2} \tag{10}$$

D. *Procedure for CFCM*

```
1. Given: subsets of patterns $X_1, X_2, \ldots\ldots, X_p$.
2. Select: distance function, number of clusters (c), termination condition, and collaboration coefficient $\beta[l,m]$.
3. Compute: initiate randomly all partition matrices $U[1], U[2], \ldots U[P]$
   -Phase I
      For each data
      Repeat
         Compute prototype { $V_j[l], j = 1,2,\ldots,C$ and partition matrices $U[l]$ for all subsets of patterns}
      Until a termination condition has been satisfied
      Communicate cluster prototype from each data site to all others;
         For each subsets of patterns $X[l], l = 1,2,\ldots,P$
         Compute, the induced partition matrices based on (2) with local data and cluster prototype from each data sites.
   -Phase II
      Repeat
         For the matrix of collaborative links $\beta[l,m]$.
         Compute, prototype $V_j[l]$ and partition matrices $U[l]$ by using (7) and (10).
      Until a termination condition has been satisfied
```

E. *Procedure for Proposed Algoritm*

1) Problem

Taking direct subtraction between $u_{ik}[l] - \tilde{u}_{ik}[l/m]$, may lose the meaning of difference between two membership degrees $u_{ik}[l]$, $\tilde{u}_{ik}[l/m]$ under different partition matrices of one pattern X_k to the same cluster. If the rows order of one matrix changes, the subtraction between two matrices will change too. The cluster describes by *k*-th row $V_k[l]$ in $u_{ik}[l]$ may be different that describe by the *k*-th row in $V_k[m]$ in $\tilde{u}_{ik}[l/m]$. In this case, taking direct subtraction between two matrices $u_{ik}[l]$ and $\tilde{u}_{ik}[l/m]$ is not a good idea.

2) Solution

Find a constructive approach of the preprocessing in order to rearrange the rows order of $u_{ik}[l]$ corresponding to the rows order of $\tilde{u}_{ik}[l/m]$ in a rational way. The match rows pair is determined by using (11).

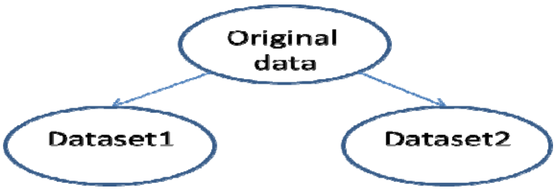

Fig. 4. An original data was divided to two different data sets

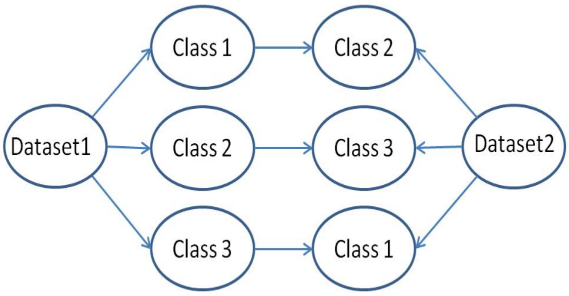

Fig. 5. Representation of classes for dataset1 and dataset2 after FCM

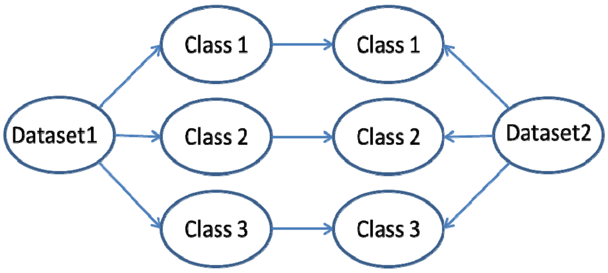

Fig.6.Representation of classes for dataset1 and dataset2 after preprocessing

$$r = \arg \min_{j=1,2..,c} \sum_{i=1}^{n} (V_{ki}[l] - V_{ji}[m])^2 \qquad (11)$$

The k-th row of $V[l]$ and the r-th row of $V[m]$ are considered to be matched row pair *(* $k = 1,2,\ldots,c$ *)*. Where n is a number of features. Similarly, update this value with $u_{ik}[l]$ and $\tilde{u}_{ik}[l/m]$.

3) Discussion

Fig. 4 shows a general way to divide a dataset into two equally different data sites. Fig. 5 shows the representation of classes for dataset1 and dataset2 after FCM. Here we can easily visualize how rows pair are mismatched and this mismatch problem leads our system in a different direction and gives a wrong sense of data analyzing. So, in order to solve this problem, this paper introduced a preprocessing process as discussed above and the benefit has been shown in Fig. 6.

4) Algorithm

Based on the above discussions and the results, we add one more phase called phase II for preprocessing and present the refined algorithm as follows:

```
1. Given: subsets of patterns $X_1, X_2, \ldots\ldots, X_p$.
2. Select: distance function, number of clusters ($c$), termination condition, and collaboration coefficient $\beta[l,m]$.
3. Compute: initiate randomly all partition matrices $U[1], U[2], \ldots U[P]$
   -Phase I
      For each data
      Repeat
         Compute prototype { $V_j[l], j = 1,2,\ldots,C$ and partition matrices $U[l]$ for all subsets of patterns}
      Until a termination condition has been satisfied
      Communicate cluster prototype from each data site to all others;
         For each subsets of patterns $X[l], l = 1,2,\ldots,P$
         Compute, the induced partition matrices base on (2) with local data and cluster prototype from each data sites.
   -Phase II
      Choose an approach for the preprocessing on cluster prototype and its corresponding partition matrices to adjust row order.
   -Phase III
      Repeat
         For the matrix of collaborative links $\beta[l,m]$.
         Compute, prototype $V_j[l]$ and partition matrices $U[l]$ by using (7) and (10).
      Until a termination condition has been satisfied
```

III. Experimental Results

A. Experiment with Iris Data Set

The Iris data [9] from UCI repository contains 3 classes of 50 instances, each with 4 attributes, where each class refers to a type of iris plant. We divided this data set into 2 equal data sets called datset1 and dataset2. Each dataset contains total 75 instances of belonging to 3 different classes with 4 attributes of each instance.

B. Experiment with Chaotic Time Series Data Set

The prediction of a chaotic time series (Mackey Glass) data has been applied to the proposed system model. This time series has been commonly used in [10-12]. MG data contains 1000 patterns and 5 attributes with 7 different classes. We divided this data set into 2 equal data sets called datset1 and dataset2, each dataset contains total 500 instances of 7 different classes with 5 attributes.

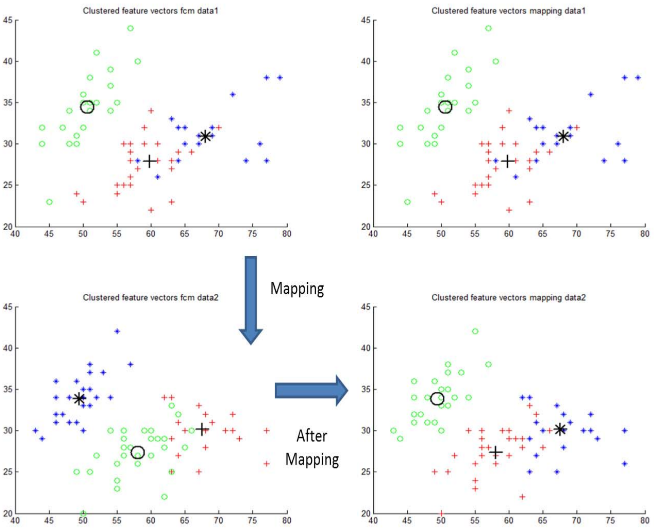

Fig. 7 Clustered feature vectors of dataset1 and dataset2

TABLE I
SIMULATION RESULT FOR IRIS DATA

β	Δ1	Δ2	Δ3
0.01	7.1009	7.0400	**3.3361**
0.03	7.6332	7.4077	**2.6330**
0.05	6.9310	6.5458	**2.6097**
0.07	6.3358	5.8692	**3.0929**
0.09	6.2483	5.6086	**2.8270**

TABLE II
SIMULATION RESULT FOR MG DATA

β	Δ1	Δ2	Δ3
0.01	14.3439	13.7881	**5.3519**
0.03	14.9255	13.8120	**5.3228**
0.05	14.8824	12.9209	**5.2698**
0.07	14.3437	12.3610	**5.2038**
0.09	13.7261	10.7127	**5.2074**

C. Discussion on Results

In Fig. 7, the first plot of row one and row 2 are clustered feature vectors of data1 and data2 respectively. As we can see in this plot, the first class of dataset1 matches with the second class of dataset2, the second class of dataset1 matches with the third class of dataset2 and the third class of dataset1 matches with the first class of dataset2. If we look at the second plot of each row of Fig. 7, this plot shows the effect of centroid mapping for prototype and rows order mapping with partition matrix through preprocessing. Now we can easily take the difference between rows of two data sites and easily do mapping between them.

Let us, consider $\Delta = \left\| u_{ik}[l] - \tilde{u}_{ik}[l/m] \right\|$, to express the degree of approximation of $u_{ik}[l]$ and $\tilde{u}_{ik}[l/m]$. In other words, Δ is a consistent analysis and it indicates the structural differences between partition matrices $u_{ik}[l]$ and $\tilde{u}_{ik}[l/m]$. The smaller Δ, the more similarity between $u_{ik}[l]$ and $\tilde{u}_{ik}[l/m]$. Table I and II show the simulation results for IRIS data and MG data respectively. Δ1 evaluates the similarity of partition matrices, before and after collaboration, Δ2 shows the similarity of partition matrices, before and after collaboration without preprocessing, and Δ3 finds the similarity of partition matrices, before and after collaboration with preprocessing.

IV. CONCLUSIONS

This paper presents the importance of preprocessing in collaborative clustering and highlights the performance of the system. Preprocessing helps and guides the collaborative phase to do collaboration in a perfect and rational way. Without preprocessing the subtraction between matrices $u_{ik}[l] - \tilde{u}_{ik}[l/m]$ misleads and violates the definition of matrix subtraction. In general, a matrix subtraction between two similar properties holding matrices mean the *i*-th row of matrix-1 should belong the *i*-th row of matrix-2 and this property was not fulfilled by methods proposed in the literature, so by doing preprocessing, the proposed method in this paper, solves this problem and keep the rationality of the system. In the future, we would apply some optimization algorithm like particle swarm optimization (PSO), differential evolutions (DE) etc. to tune the parameter in order to keep the higher performance of system and better understanding of the unknown future of unknown datasets.

ACKNOWLEDGMENT

This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan National Science Council I-RiCE program under grant number: NSC 102-2911-I-009 -101, in part by NSC-NSC102-2220-E-009-041, in part by the aiming for the Top University Plan, under contract 103W963.

REFERENCES

[1] J.C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," Plenum Press, New York, 1981.

[2] J.C. Bezdek, R. Ehrlich, W. Full, "FCM: the fuzzy C-means clustering algorithm," Computers and Geosciences.

[3] W. Pedrycz , "Knowledge-based clustering: from data to information granules," A John Wiley&Sons, Inc., Publication, 2005.

[4] W. Pedrycz , "Collaborative fuzzy clustering," Pattern Recognition Letters, Vol. 23, No. 14, pp. 1675–1686, 2002.

[5] W. Pedrycz and P. Rai, "Collaborative Fuzzy Clustering with the use of Fuzzy C-Means and its Quantification," Fuzzy Sets and Systems, DOI 10.1016/j. fuss. 2007.12.030, 2008.

[6] L. F.S. Coletta, L. Vendramin, E.R. Hruschka, R. J. B. Campello and W. Pedrycz, "Collaborative Fuzzy Clustering Algorithms: Some Refinements and Design Guidelines," IEEE Trans. On Fuzzy Syst., vol. 20, no. 3, jun. 2012.

[7] M. Prasad, C.T Lin, C.T Yang and A. Saxena, "Vertical Collaborative Fuzzy C-Means for Multiple EEG Data Sets," Springer Lecture Notes in Computer Science, Vol. 8102, pp. 246-257, 2013.

[8] C.T Lin, M. Prasad, and A. J.Y Chang, "Designing Mamdani Type Fuzzy Rule Using a Collaborative FCM Scheme," 2013 International Conference on Fuzzy Theory and Its Application.

[9] http://archive.ics.uci.edu/ml/datasets/Iris.

[10] J. R. Castro, O. Castillo, P. Melin, and A. Rodríguez-Díaz, "A hybrid learning algorithm for a class of interval type-2 fuzzy neural networks," Inform. Sci., vol. 179, no. 13, pp. 2175–2193, Jun. 2009.

[11] S. Hengjie, M. Chunyan, S. Zhiqi, M. Yuan, and B. S. Lee, "A fuzzy neural network with fuzzy impact grades," Neurocomput., vol. 72, nos. 13–15, pp. 3098–3122, Aug. 2009.

[12] D. Kim and C. Kim, "Forecasting time series with genetic fuzzy predictor ensembles," IEEE Trans. on Fuzzy Syst., vol. 5, no. 4, pp. 523–535, Nov. 1997.