2012 International Conference on Software and Computer Applications (ICSCA 2012)
IPCSIT vol. 41 (2012) © (2012) IACSIT Press, Singapore

Comparative Analysis of Fuzzy Clustering Algorithm for Industrial Process Monitoring

Kiran Jyoti [1] and Dr. Satyaveer Singh[2]

[1]Assistant Professor, Department of IT, GNDEC, Ludhiana.
kiranjyotibains@yahoo.com
[2]Assistant Professor, Department of Mathematics,
JJT University, Jhunjhunu, Rajasthan
s.angasar@rediffmail.com

Abstract. In modern day collection of information from different sources, classification of information according to their property, traits and arranging the information according to the sub group is very important in nature. The whole task of collection, classification and clustering (arranging the sub group of data) is very challenging in nature. Different researchers are working to find a better way to classify data, recognize the data and cluster the data. The data collection, classification and clustering is of prime importance of different industry. In industrial plant, where there are thousands of sensors and transducer to provide online streaming of process variable data, data collection, data processing is a very challenging job. The data processing done using these steps in the industry helps the management to see the overall performance of the plant and act on the drawbacks of the system to fine tune the system. In this paper an industrial process monitoring application is taken in to consideration for implementation of different fuzzy based clustering algorithms. A PSO (Particle Swarm optimization method) and fuzzy based clustering algorithm is proposed in this paper and this proposed hybrid algorithm is compared with other existing fuzzy algorithms by taking two cluster validity indices. Results shows proposed hybrid FPSO can cluster the subgroup of data in an efficient manner.

Keywords: FCM, PSO, Clustering, Fuzzy

1. Introduction

Mostly the industrial plants are non linear and multiple input-multiple output MIMO in nature. In industrial plant there are thousands of sensors and transducers which sense different physical parameter in the field and provide the value of parameters (data) to the controller and the controller subsequently forwards all the data for the control room for record keeping and statistical process control application. In the control room all the data coming from the sensor are fused together which is called as multi sensor data fusion. After the multi sensor data fusion, different pattern recognition algorithm is used to recognize different form of data and pattern classification is used for classification of different sub groups of data. Data clustering algorithms are used for clustering the same group of data which can further optimized by different stochastic global optimization techniques and evolutionary algorithm can be used. Some of the example of these kind of algorithms are particle swarm optimization, genetic algorithm, stimulated annealing, ant colony optimization, genetic programming.

Different clustering algorithm can be used for clustering purpose. Some of the data clustering algorithms which can be used are fuzzy C means algorithm, fuzzy K means algorithm. Fuzzy C means algorithm is very popular as it is easy to use, straight forward and very efficient. However fuzzy C means algorithm is very sensitive to initialization and gets trapped very easily in local optima. The said clustering algorithms can be optimized using PSO algorithm which is a stochastic global optimization tool. A hybrid fuzzy clustering

method based on fuzzy C means algorithm and fuzzy based particle swarm optimization method can be used for better clustering algorithm.

2. Fuzzy Algorithms

2.1. Fuzzy c-means clustering

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 and improved by Bezdek in 1981 and it is frequently used in pattern recognition. In fuzzy clustering, one object can be clustered in more than one cluster according to the degree of membership function. Consider a set of n vectors $X \rightarrow \{x_1, x_2, x_3 \ldots\ldots x_n\}$ has to be clustered in to $C \rightarrow \{c_1, c_2, c_3 \ldots\ldots c_k\}$. $d(x,C_i)$ denote the similarity between object x and cluster C_i. where $2 \leq c \leq n$. [2,4,13] The FCM algorithm is described by taking a fuzzy membership matrix called a fuzzy partition represents fuzzy clustering of the object. Let M_{fc}= c×n fuzzy partition matrix where

$$U=M_{fc} = \begin{pmatrix} .60 & .50 & \ldots\ldots \\ .30 & .25 & \ldots\ldots \\ .10 & .25 & \ldots\ldots \end{pmatrix} \quad (1)$$

$U_{ik} \in [0,1]$, $\forall$ i,k constraints in M_{fc} are sum of centers is always and number of datasets are always between (0,1).

$\sum_{i=1}^{c} U_{ik} = 1 \ \forall$ k sum of partition matrix is always $10 < \sum_{k=1}^{n} U_{ik} < n, \ \forall$ i no of clusters should be less than no of vectors n.The main aim of fuzzy clustering is to minimize following objective function:

$$J_m (U,P) = \sum_{k=1}^{n} \sum_{i=1}^{c} (U_{ik})^m \left\| x_k - p_i \right\|^2 \quad (2)$$

Where m $\in (1, +\infty)$ ***is*** the ***weighing exponent P=(P1, P2, P3…..Pi...,Pc) is the vector of clustering***

Fuzzy C means tries to minimize J_m by iteratively updating the partition matrix using following equation

$$P_i = \sum_{k=1}^{n} (U_{ik})^{m} x_k \sum_{k=1}^{n} (U_{ik})^{m} \left(\sum_{k=1}^{n} (U_{ik}) \right)^{-m} \quad (3)$$

$$\& \quad U_{ik}^{(b)} = \left\{ \sum_{j=1}^{c} \left[\frac{di_k^{(b)}}{d_{jk}^{(b)}} \right]^{2(m-1)} \right\}^{-1} \text{ for } b^{th} \text{ iteration} \quad (4)$$

This algorithm stops if terminating condition *is met i.e.* $\left\| P^{(b)} - P^{(b+1)} \right\| < \varepsilon$ *where ε is improvement factor and in this value is taken* $1e^{-4}$

2.1.1 Algorithm

Step 1. *Define data number and cluster number.*
Step 2. *A fuzzy partition matrix U is defined as in equation (1)*
Step 3. *Define an objective function for creating clusters as in (2)*
Step 4*.Create initial random clusters.*
Step 5. *After clusters based on Fuzzy membership functions have been created, find the distance between each data point and the cluster head.*
Step 6. *Calculate distance as in step 5.*
Step 7. *Compare the distance with previous iteration. If distance is lesser then terminate else go for updating distance as in equation (4).*

2.2. Gustafson-Kessel algorithm

The Gustafson-Kessel algorithm associates each cluster with both a point and a matrix, respectively representing the cluster centre and its covariance. Whereas the original fuzzy c-means make the implicit hypothesis that clusters are spherical, the Gustafson-Kessel algorithm is not subject to this constraint and can identify ellipsoidal clusters.[16]

2.2.1 Algorithm

Step1.. *Initialization: Calculate the initial number of clusters c and the corresponding matrices; initial centroids matrix V, initial fuzzy partition matrix U and initial cluster covariance matrix F. Choose termination tolerance tol.*

Step2. *Do While max (max ($U_t - U_{t+1}$)) > tol*

Repeat for every data point Xk

i. *Calculate Mahalanobis-like distances d_{ik} , i=1....c, using:*

$$d_{ik} = \sqrt{\{(X_i - V_i)[\left(\det(F_i)\right)^{\frac{1}{n}} F_i^{-1}](X_k - V_i)^T\}}$$

ii. *Determine the closest cluster p through minimal distance, by* $p = \arg\min(d_{ik}), \forall i = 1, ..., c$

iii. *Update the center vi and matrix Fi*

$$v_i = \frac{\sum_{k=1}^{N} (U_{ik})^m X_k}{\sum_{k=1}^{N} (U_{ik})^m} \quad \text{and} \quad F_i = \frac{\sum_{k=1}^{N} (U)^m (X_k - V_i)^T (X_k - V_i)}{\sum_{k=1}^{N} (U_{ik})^m}$$

Recalculate the partition matrix U where . $U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)}}$

End

Step3. *When the termination condition is met, the partitions are obtained through last U and V.*

3. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is based on the collective motion of a flock of particles: the particle swarm. In the simplest and original version of PSO, each member of the particle swarm is moved through a problem space by two elastic forces. One attracts it with random magnitude to the best location so far encountered by the particle. [3, 5, 10] The other attracts it with random magnitude to the best location encountered by any member of the swarm. PSO consists of a swarm of particles and each particle flies through the multi dimensional search space with a velocity, which is constantly updated by the particle‘s previous best performance and by the previous best performance of the particle‘s neighbors. PSO can be easily implemented and is computationally inexpensive in terms of both memory requirements and CPU speed. The position and velocity of each particle are updated at each time step (possibly with the maximum velocity being bounded to maintain stability) until the swarm as a whole converges to an optimum. In clustering based on PSO, PSO is generally applied on centroid of clusters to optimize them. The basic algorithm is as follows:

N_d is dimension of data, c is no. of clusters, z_p is p^{th} data vector, m_j is j^{th} cluster centroid vector, n_j is no of data vector in j^{th} cluster

3.1. Algorithm

Step1.*Initialize each particle to contain c randomly selected cluster centroid.*

Step2. *For iterations i to max*

(a) *For each particle*

(b) *For each vector z_p*

(i) *Calculate Euclidean distance of each particle to each centroid*

i.e. d(z_p, m_{ij}) i.e. $\|z_p - m_{ij}\|$

(ii) *Assign the data vector (element) zp to one cluster such that its distance from its centroid is minimum from other cluster’s centroids i.e. d (z_p, m_{ij})= min* $\forall$ *c =1, ...M_c* $\{d(z_p, m_{ic})\}$

(iii) *Calculate fitness using fitness function*

$$J_e = \frac{\sum_{j=1}^{c} \left[\sum_{\forall z_p \epsilon c_{ij}} d \frac{(z_p, m_j)}{|c_{ij}|} \right]}{N_c}$$ *where* $|c_{ij}|$ *is the no. of data vectors belonging to cluster c_{ij}*

(c) *Calculate and update* p_{best} *and* g_{best}

(d) *Update one cluster centroids using*

$$v_{i,k}(b+1) = wv_{i,k}(t) + c_1 r_1\left(p_{best,k}^{(t)} - x_{i,k}(t)\right) + c_2 r_2\left(g_{best,k}^{(t)} - x_{i,k}(t)\right) \tag{5}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{6}$$

here w is inertia constant; c_1 c_2 *acceleration constant ;* r_1 r_2 *are random no's,* x_i *here represents centroid vectors not particles i.e. data vectors.*

Step3. *The final centroid is in the* p_{best} *after the max iterations are over.*

4. Proposed Hybrid algorithm

FCM is one of the most commonly used fuzzy clustering techniques for different degree estimation problems. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. FCM restriction is the clusters number which must be known a priori. FCM employs fuzzy partitioning such that a data point can belong to several groups with the degree of membership grades between 0 and 1 and the membership matrix U is constructed of elements that have value between 0 and 1.The aim of FCM is to find cluster centers that minimize a dissimilarity function. U is the membership matrix, is randomly initialized. In the fuzzy clustering, a single particle represents a cluster center vector, in which each particle *Pl* is constructed as follows $P_l = (V_1, V_2, -----V_i---V_c)$ where i represents number of clusters. V_i is the vector of c^{th} cluster center. Vi = (Vi1, Vi2, -------- Vid) where 1, 2,…, d are dimensions of cluster center vectors. Therefore, a swarm represents a number of candidates clustering for the current data vector. In this proposed hybrid method the objective function of FCM is modified and after modification it becomes the fitness function like that of PSO i.e.

$$f(x) = \frac{K}{J_m + \xi} \tag{7}$$

where K and ξ *are constants.*

Each point or data vector belongs to every various cluster by different membership function, thus, a fuzzy membership is assigned to each point or data vector. For the purpose of this algorithm, following notations are defined:

n : number of data vector, C : number of cluster center, $V_l^{(t)}$ *: position of particle i in stage t*

4.1. Algorithm

Step 1. *Initialize the fuzzy partitions and centers where centers* $U = \dfrac{\sum_{k=1}^{n} (U_{ik})^m x_k}{\sum_{k=1}^{n} (U_{ik})^m}$

Step2. *Calculate distances by equation i.e* $U_{ik} = \dfrac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}$ *and calculate fuzzy membership function.*

Step 3. *Initialize swarm (according to cluster centroids) ,* p_{best} *&* g_{best}

Step 4. *For iteration t to* t_{max} *calculate the fitness of each centroid particle using equation (7)where* J_m *is given in equation 7*

Step 5. *Update* p_{best} *and* g_{best}

Step 6. *Calculate the new centroids using PSO by equations (5) and (6) for position and velocity*

$$v(t+1) = wv(t) + c_1 r_1 (p_{best} - x(t)) + c_2 r_2 (g_{best} - x(t))$$

$$x(t+1) = x(t) + v(t+1)$$

Step 7. *Go to step 4 if terminating condition has not met else stop*

5. Validity Indices

Cluster validity refers to the problem whether a given fuzzy partition fits to the data at all. Here in this research for validating the proposed algorithm two validity indices are used. One is Partition Coefficient (PC) and other is Partition Entropy (PE).[11,15]

Partition coefficient: An index which measures the fuzziness of the partition but without considering the data set itself. It is a heuristic measure since it has no connection to any property of the data. The maximum values of it imply a good partition in the meaning of a least fuzzy clustering. The first validity index associated with FCM was the partition coefficient defined by

$$PC(c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij} \quad where\ 1/c \le PC(c) \le 1. \tag{8}$$

In general, we find an optimal cluster number c* by solving $\max_{2\le c\le n-1}$ PC(c) to produce a best clustering performance for the dataset X. That is we have to maximize partition coefficient.

Partition entropy: It is a measure that provides information about the membership matrix without also considering the data itself.[8,15] The minimum values imply a good partition in the meaning of a more crisp partition. The partition entropy was defined by

$$PE(c) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}\log_2 U_{ij} \quad where\ 0 \le PE(c) \le log_2 c \tag{9}$$

In general, we find an optimal c* by solving $\min_{2\le c\le n-1}$PE(c) to produce a best clustering performance for the data set X.

6. Comparison of Fuzzy clustering Algorithms

In fuzzy clustering role of validity index is very important. It helps to determine the appropriate number of clusters in the dataset. Here in this research two validity indices are taken one is partition coefficient (PC) as described in equation (9) and second is partition entropy (PE) as described in equation (10). With this optimum number of clusters can be found by max c between 2 and n-1 for PC and min c between 2 and n-1 for PE to produce best clustering performance for dataset N.

Fuzzy algorithms are compared with proposed FPSO algorithm using validity indices PC and PE for fuzziness exponent 2 and 2.5 for different dataset as shown in table 1 and table 2 respectively. It is observed from both the tables that as number of clusters increase the PC value decreases for GK as well as for FCM but it increase for proposed PSO for fuzzyfication exponent m=2 and m=2.5. But PE increases for GK ,FCM and proposed hybrid algorithm but still the values of PE is less for proposed hybrid PSO as compared to GK and FCM for m=2.5. So as compared to other fuzzy algorithms the performance of proposed hybrid PSO is better for this industrial process monitoring application and proposed method is giving best result with any fuzziness exponent value 2.5. if we talk about PC index in that case best result for c=4 and

Table 1Comarasion of fuzzy algorithms with proposed hybrid *FPSO for dataset values 2500*

M	No. of clusters	Techniques	PC	PE
m=2	2	gustafson kessel	0.7153	0.4412
		Fuzzy c-means	0.7287	0.4242
		Hybrid FPSO	**0.6673**	**0.4999**
	3	gustafson kessel	0.6202	0.6652
		Fuzzy c-means	0.5575	0.7553
		Hybrid FPSO	**1.0093**	**0.7472**
	4	gustafson kessel	0.5331	0.8790
		Fuzzy c-means	0.5007	0.9252
		Hybrid FPSO	**1.3296**	**0.9985**

m=2.5	2	gustafson kessel	0.6275	0.5518
		Fuzzy c-means	0.6341	0.5445
		Hybrid FPSO	**0.6678**	**0.4982**
	3	gustafson kessel	0.4869	0.8729
		Fuzzy c-means	0.4841	0.8750
		Hybrid FPSO	**1.0030**	**0.7496**
	4	gustafson kessel	0.3992	1.1223
		Fuzzy c-means	0.3772	1.1532
		Hybrid FPSO	**1.3358**	**0.9992**

Table 2 Comparison of fuzzy algorithms with proposed hybrid *FPSO for dataset values 10,000*

M	**No. of clusters**	**Techniques**	**PC**	**PE**
m=2	2	gustafson kessel	0.7222	0.4325
		Fuzzy c-means	0.7379	0.4124
		Hybrid FPSO	**0.6622**	**0.4985**
	3	gustafson kessel	0.6216	0.6605
		Fuzzy c-means	0.6054	0.6831
		Hybrid FPSO	**1.0009**	**0.7441**
	4	gustafson kessel	0.5482	0.8483
		Fuzzy c-means	0.5199	0.8958
		Hybrid FPSO	**1.3451**	**0.9964**
m=2.5	2	gustafson kessel	0.6318	0.5470
		Fuzzy c-means	0.6498	0.5261
		Hybrid FPSO	**0.6739**	**0.4986**
	3	gustafson kessel	0.4883	0.8700
		Fuzzy c-means	0.4932	0.8616
		Hybrid FPSO	**1.0033**	**0.7482**
	4	gustafson kessel	0.4027	1.1149
		Fuzzy c-means	0.3749	1.1561
		Hybrid FPSO	**1.3494**	**0.9955**

m=2.5. Proposed hybrid FPSO algorithm will be with no of clusters c=4 and m=2.5. Even PE will be less comparatively other fuzzy algorithms. If we talk about PE index in that case best results of proposed hybrid FPSO algorithm will be with c=2 and m=2.5.

7. Conclusion and Future Scope

In industrial plants online streaming of process data, data collection and data processing is very challenging job so different data clustering algorithms have to be used. In this paper different fuzzy based clustering techniques for fault detection in process plant monitoring are implemented. A large dataset is used for process plant monitoring. Gustafson-Kessel, FCM clustering algorithms are applied and implemented on this process monitoring dataset then a hybrid PSO (FPSO) is proposed. The quality of clustering is validated and compared with this proposed algorithms. Taking two validity indices partition coefficient (PC) and

partition entropy (PE). Proposed hybrid FPSO is giving high values of PC for fuzzyfication exponent 2.5 with lowest value of PE at fuzzyfication exponent 2.5 than other fuzzy algorithms.

Future Scope: In this paper all the algorithms are implemented on 2-Dimessional data it can further be tried on multidimensional dataset.

8. Acknowledgements

Authors thank to JJT University and GNDEC for their support and help for carrying out this research. We also thank the reviewers, the referees and the editors, for suggestions, comments, help with experiments and data**;**

9. References

[1] Anil K Jain, "Data Clustering: 50 Years Beyond K Means," Pattern Recognition Letters, (2010) 31, pp. 651-666

[2] Lin Zhu, Fu-Lai Chung, And Shitong Wang, "Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, (2009) vol. 39, No. 3, Page No 578-584.

[3] Ying-Xin Liao, Jin-Hua She, and Min Wu, "Integrated Hybrid PSO and Fuzzy-NN Decoupling Control for Temperature of Reheating Furnace," IEEE Transactions On Industrial Electronics, (2009) vol. 56, no. 7, pp. 2704-2714

[4] Derek T. Anderson, Robert H. Luke, and James M. Keller, "Seedup of Fuzzy Clustering Through Stream Processing on Graphics Processing Units," IEEE Transactions on Fuzzy Systems, (2008) vol. 16, no. 4, pp. 1101-1106

[5] K Premalatha and A M Natarajan, "A New Approach for Data Clustering Based on PSO with Local Search," Computer and Information Science, (2008) vol. 1, no. 4, pp. 139-145

[6] Osama Abu Abbas, "Comparisons Between Data Clustering Algorithms," The International Arab Journal of Information Technology, (2008) vol. 5, no. 3, pp. 320-325

[7] T. Niknam, M. Nayeripour and B.Bahmani Firouzi, "Application of a New Hybrid Optimization Algorithm for Cluster Analysis," World Academy of Science, Engineering and Technology, (2008) vol. 46,pp. 589-594

[8] C Lionberger and M Cromaz, "Control of Acquisition and Cluster Based Online Processing of Gretina Data," Proceedings of ICALEPCS 07, (2007) pp. 93-95

[9] Gursewak S. Brar, Yadwinder S Brar and Yaduvir Singh, "Implementation and Comparison of Contemporary Data Clustering techniques for Multi Compressor System: A Case Study," WSEAS Transactions on Systems and Control, (2007) no 9, issue 2, pp. 442-449

[10] Sherin M Youssef, Mohamed Rizk and Mohemad El-Sherif, "Dynamically Adaptive Data Clustering Using Intelligent Swarm-like Agents," International Journal of Mathematics and Computers in Simulation, (2007) vol. 1, issue 2, pp. 108-118

[11] WeinaWang ,Yunjie Zhang," On fuzzy cluster validity indices", Science Direct journal on Fuzzy Sets and Systems (2007) pp. 2098-2116

[12] Zhe Song and Andrew Kusiak, "Constraint Based Control of Boiler Efficiency: A Data Mining Approach," IEEE Transactions on Industrial Informatics, (2007) vol. 3, no. 1,pp. 73-83

[13] Skrjanc I., "Fuzzy Model Based Detection of Sensor Faults in Waste Water Treatment Plant," in Proceedings of 5th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, (2006) pp. 195-199

[14] Yi-Tung Kao, Erwie Zahara and I-Wei Kao, (2006) "A Hybridized Approach to Data Clustering," in Proceedings of the 7th Asia Pacific Industrial Engineering and Management Systems Conference ,pp. 497-504

[15] Kuo-Lung Wu a, Miin-Shen Yang (2005)," A cluster validity index for fuzzy clustering", ", Science Direct journal on Fuzzy Sets and Systems pp. 1275-1291

[16] Raghu Krishnapuram and Jongwoo Kim,(1999) "A Note on the Gustafson–Kessel and Adaptive Fuzzy Clustering Algorithms" IEEE Transactions on fuzzy systems, vol. 7, no. 4, pp 453-461