Data Clustering Approach to Industrial Process Monitoring, Fault Detection and Isolation

Kiran Jyoti
Department of CSE and IT
GNDEC, Ludhiana, Punjab, India

Dr. Satyaveer Singh
Department of Mathematics
JJT University, Jhunjunu, Rajasthan

ABSTRACT

In this paper proposes different conventional and fuzzy based clustering techniques for fault detection and isolation in process plant monitoring. Process plant monitoring is very important aspect to improve productiveness and efficiency of the product and plant. This paper takes a case study of plant data and implements K means algorithm and fuzzy C means algorithm to cluster the relevant data. This paper also discusses the comparison for K means algorithm and fuzzy C means algorithm.

General Terms: Pattern Recognition, Data Clustering.

Keywords: Conventional Clustering, Fuzzy Based Clustering, Fault Detection and Isolation

1. INTRODUCTION

Clustering is a typical unsupervised learning technique for grouping similar data points. A clustering algorithm assigns a large number of data points to a smaller number of groups such that data points in the same group share the same properties while, in different groups, they are dissimilar. Clustering has many applications, including part family formation for group technology, image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis.

One of the best known and most popular clustering algorithms is the k-means algorithm. The algorithm is efficient at clustering large data sets because its computational complexity only grows linearly with the number of data points. However, the algorithm may converge to solutions that are not optimal.

Many clustering methods have been proposed. They can be broadly classified into four categories: partitioning methods, hierarchical methods, density-based methods and grid-based methods. Other clustering techniques that do not fit in these categories have also been developed. These are fuzzy clustering, artificial neural networks and genetic algorithms.

2. INDUSTRIAL PROCESS MONITORING, FAULT DETECTION AND ISOLATION

Monitoring is a continuous real-time task of determining the conditions of a physical system, by recording information, recognizing and indicating anomalies in the behavior. In other words, the purpose of the monitoring is to indicate whether a process has deviated from its acceptable state, and if it has, why. The deviations are called process faults. Observation of the faults is known as fault detection, which is followed by fault isolation, determination of the location and the type of the fault. Fault Detection and Isolation (FDI) – also known by a common name fault diagnosis – can be carried out in many ways. The three logical parts of any FDI scheme are namely detection, decision and isolation, may be partially integrated. Fault detection takes as input the current values of the process measurements and produces one or more fault indicator signals, which are often called residuals. After the detection phase there is an inference mechanism which takes the fault indicator(s) as input and decides whether a fault has occurred or not. Detection of a fault is followed by an isolation phase which carries out identification of the fault.

Fault detection methods are divided into two categories: first principles process models and models of process data. In the former approach, physical structure and a priori known relationships between variables of a process form the basis for the construction of the model and observed data are not required. In the latter case, the structure of the model is generic or depends on the data and the model is based on observed data produced by the sensors of the process. However, in practice the line between these two categories is not sharp: measurement data may be used in construction of a first principles model and, correspondingly, a priori knowledge of a process can be used in the construction of a model using process data. Use of data-driven models instead of the first principles models is justified if construction of an accurate first principles model is

(1) Impossible due to unknown process phenomena

(2) Computationally infeasible because of complexity of the process. In addition, if the modeling using first principles would be possible but laborious, use of a data-based model may still be reasonable choice if the process is modified often: a model based on data is typically easier to update than a first principles model.

3. RELATED WORK

Scott C Newton et.al in their paper proposed a hybrid adaptive fuzzy leader clustering technique implemented in ART-I like structure to cluster speech, image and medical data [1].

A K Jain et.al in his seminal work has given a detail overview of data clustering, application of fuzzy logic, artificial neural network, and genetic algorithm in clustering algorithm. Many of his review papers on data clustering are published in well known journals and accepted world wide [3, 4, 17].

Y M Sebzalli et.al, has proposed two techniques like principal component analysis (PCA) and fuzzy C means clustering to identify and develop operational strategy for manufacture of desired product in process industry. This research paper takes a case study of fluid catalytic cracking process used in refinery industry. The authors analyzed the problem by collecting three hundred data from the process site and applying principal component analysis and fuzzy c means clustering algorithm in the datasets [5].

Timo Ahvenlampi et.al, studied the controllability of kappa number in two cooking application. Kappa number is the quality measure of pulp cooking method. The authors investigated the clustering and fault diagnosis approach of cooking system [6].

Young-Hak Lee et.al has proposed an adaptive monitoring technique of real time industrial process to classify and distinguish operational changes. The proposed method extracts process knowledge and classifies process state changes. The case study taken by the authors is a refinery fired heater [7].

Skrjanc I has presented in his research paper a method of sensor fault detection in waste water treatment plant using Gustafson-Kessel fuzzy clustering algorithm. Different measurements like influent ammonia concentration, dissolved oxygen concentration in aerobic reactors are measured and analyzed [8].

C Lionberger et.al has proposed a novel method of online acquisition and clustering of GRETINA (Gamma Ray Energy Tracking In-beam Nuclear Array) which is an array of 28 36-segment germanium crystals [10].

Zhe Song et.al has proposed a data mining approach to develop a model for optimizing the efficiency of an electric utility boiler. The industrial boiler generates real time data used for clustering. The clustering algorithm learns and generates new knowledge used to update the control signature database. Based on the real-time boiler status, the optimization algorithm searches the control signature database for an optimal centroid controlling the process. Thus, the boiler performance is improved [11].

N Sujatha et.al has proposed in her research paper an innovative way to find out the web usage pattern by implementing modified K means algorithms and optimizing the cluster quality by using genetic algorithm based refinement algorithm. The modified K means algorithm and refinement algorithm based on genetic algorithm is applied in web access log collected from internet traffic archive (ITA) [14].

Osama Abu Abbas in his research paper compared different conventional and intelligent clustering algorithms according to the size of data sets, number of clusters, and type of datasets to find out the performance of the clustering algorithm, quality of clustering and accuracy of the clustering [15].

K Premalatha et.al in her research paper applied swarm intelligence technique like particle swarm intelligence (PSO) in cluster analysis. With application of PSO the optimal shape of cluster can be found out [16].

V Kavitha et.al, has given a literature review of clustering of time series data stream. A time series data is being generated at a unique speed from almost every application domain. These types of dataset are special type of dataset which has temporal ordering [18].

Hesam Izakian and Ajith Abraham proposed a hybrid fuzzy C means algorithm which implements fuzzy particle swarm optimization algorithm in fuzzy C means algorithm. This proposed hybrid algorithm make use of merit of both the algorithms and finds out optimal cluster structures [19].

S Kalyani and K S Swarup proposed a supervised fuzzy C means algorithm for security assessment and classification of power system. The proposed algorithm is tested on 39 bus New England and IEEE 57 bus systems. The classification results of supervised fuzzy C means algorithm is tested with method of least squared and multi layered perceptron classifiers [20].

Xian-Xia Zhang et.al, proposed a novel sensor placement technique by utilizing main feature of spatial distribution [21].

Mika Liukkonen et.al in their research paper described the dependencies between process variables and the concentrations of gaseous emission components. They also created multivariate nonlinear models describing their formation in the process. A process model was created using self organizing map and was clustered using K means algorithm for determination of subsets [22].

Vasil Simeonov et.al, has proposed a novel method of water quality assessment of high mountain lakes in Pirin Mountain in Bulgaria by application of cluster analysis and principal component analysis. The authors have also studied the classification of dataset by using self organizing map [23].

Ibrahim Masood and Adnan Hassan proposed an ANN based control chart pattern recognition system for process plant monitoring and control. The feature based and wavelet denoise method is used for input representation [24].

4. PROBLEM FORMULATION

4.1 K Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move anymore. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \qquad (1)$$

where $\left\| x_i^j - c_j \right\|^2$ is a chosen distance measure between a data point x_i^j and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centers.

4.2 Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

4.3 Need of Fuzzy Clustering

In traditional clustering algorithm, one object is assigned in to only one cluster. This is valid till the clusters are disjoint and separate. But if the clusters are touching each other or they are overlapping, then one object can belong to more than one cluster. In this case fuzzy clustering comes in to existence.

In fuzzy clustering, one object can be clustered in more than one cluster according to the degree of membership function.

Let a set of objects $X = \{x_1, x_2, x_3 ....x_n\}$ has to be clustered in to $C = \{c_1, c_2, c_3 ....c_k\}$. $\delta(x, C_i)$ denote the similarity between object x and cluster C_i. The membership function for object x and cluster C_i is represented by the following equation

$$fc_i(x) = \frac{P_i \delta(x, C_i)}{\sum_{k=1}^{K} P_k \delta(x, C_k)} \qquad (2)$$

$P_k = \frac{n_k}{n}$ is the relative size of cluster C_k. This membership function is non negative. Membership function can also be expressed in terms of Euclidian distance. This is represented in following equation

$$fc_k(x) = \frac{1 - \left(\frac{1}{\beta}\right) d \quad x, m^k}{K - \left(\frac{1}{\beta}\right) \sum_j d \quad x, m^j} \qquad (3)$$

$d \quad x, m^k$ represent the Euclidian distance between vector x and centroid m^k of cluster C_k. β denotes the belongingness.

4.4 Fuzzy based Clustering

Initially a fuzzy partition matrix U is generated that is of size N x c, where c is number of clusters and N is total number of feature vectors. Subject to the constraint that $\sum_{j=1}^{c} U_{ij} = 1$ (4)

i = {1,2,3 -------- N}

Calculation of fuzzy centers

The fuzzy centers are calculated using the partition matrix generated

$$C_j = \frac{\sum_{i=1}^{N} U_{ij}^m x_i}{\sum_{i=1}^{N} U_{ij}^m} \qquad (5)$$

where *m* ≥ 1 is a fuzzification exponent. The larger the value of *m* the fuzzier the solution will be. This indicates the number of iterations that is required for clustering. xi is ith feature vector. The value of *i* ranges from 1 to *N* (total number of templates in the database).

Updating membership and cluster centers

FCM is an iteration loop. The method of clustering is based on minimization of the objective function defined by

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} U_{ij}^m \left\| x_i - c_j \right\|^2 \qquad (6)$$

Uij describes the degree of member of feature set (*xi*) with cluster *cj*. ||*|| represents norm between *xi* and cluster center *cj* given by $\left\| x_i - c_j \right\|^2 = \; x_i - c_j \;^T A \; x_i - c_j$ (7)

where *A* is identity matrix for Euclidean distance used here. At every iteration the membership matrix is updated using

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_j \right\|} \right)^{\frac{2}{m-1}}} \qquad (8)$$

The revised membership matrix is used for updating the cluster centers.

The iteration will stop when

$$\max_{ij} \left| U_{ij}^{m+1} - U_{ij}^{m} \right| < \varepsilon \qquad (9)$$

where *ε* is a termination criteria. The value of *ε* ranges between 0 and 1.

Algorithm for fuzzy C means algorithm

1. Fix 1 < m < ∞, initial partition matrix U (N x c) and termination criteria
2. Calculate fuzzy cluster centers
3. Update membership matrix

4. Calculate change in membership function $\Delta = \left\| U^{m+1} - U^{m} \right\| = \max_{ij} \left| U_{ij}^{m+1} - U_{ij}^{m} \right|$

   If $\Delta \leq \varepsilon$ then set m = m+1 and go to step 2

   Or else stop

4.5 Gustafson-Kessel Clustering

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.

5. SIMULATION AND TESTING

This section takes a process plant in to consideration and implements different conventional and fuzzy based data clustering approach to cluster the process data.

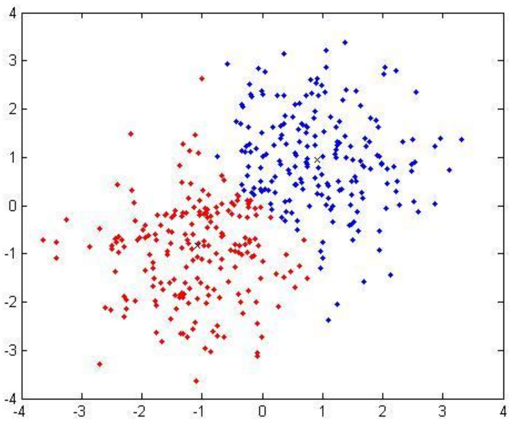
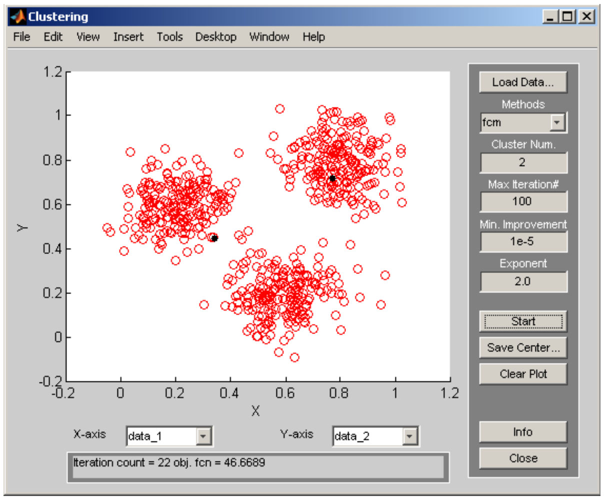


Figure 1: K Means clustering algorithm implemented in process data

Figure 1 shows the K means clustering graph implemented in process data. An arbitrary set of process data is taken and K means clustering algorithm is implemented in those data.



Methods: fcm

Cluster Num.: 2

Max Iteration#: 100

Min. Improvement: 1e-5

Exponent: 2.0

X-axis data_1

Y-axis data_2

```
Iteration count = 22 obj. fcn = 46.6689
```

Figure 2: Plot for fuzzy C means algorithm

Figure 2 shows the plot for fuzzy C means clustering.

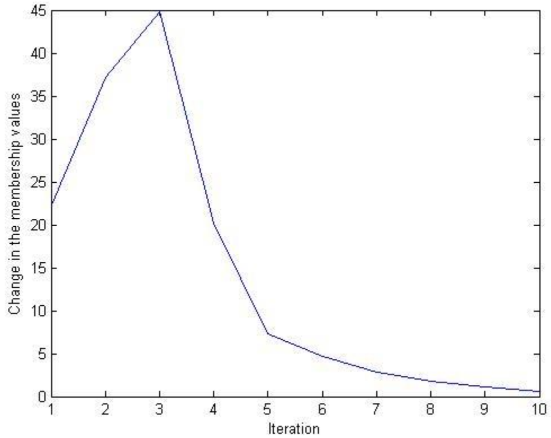


Figure 3: Plot for change in membership function with respect to iteration in fuzzy C means algorithm

Figure 3 shows the change in membership function with respect to the iteration in fuzzy C means algorithm.

6. RESULTS AND DISCUSSION

As was expected the K-means algorithm was strongly sensitive to initial points, the quality of the obtained final clusters depended strongly on the given initial set of clusters. Bins and centroid provided the best initials clusters for this algorithm. The C-Means algorithm performed really well, in all datasets the correctness obtained was comparable to the best ones achieved. This algorithm was not the faster but it was not the slower either, so the performance speed was acceptable.

7. CONCLUSION

This paper describes an efficient clustering method to improve the productivity and efficiency of process plant monitoring, fault detection and isolation. This paper also describes the related work carried out in this field. This paper takes process data and uses K means algorithm and fuzzy C means algorithm to cluster the process data. It gives a comparative study of clustering using K means algorithm and fuzzy C means algorithm. As a future scope of paper neural network like self organizing map (SOM) can be used to cluster the data. Swarm intelligence algorithms like PSO can be implemented along with fuzzy C means algorithm to find out the optimal clusters.

8. REFERENCES

[1] Scott C Newton, Surya Pemmaraju and Sunanda Mitra, “Adaptive Fuzzy Leader Clustering of Complex Data Sets in Pattern Recognition,” IEEE Transactions on Neural Network, vol. 3, no. 5, 1992, pp. 794-800

[2] E L Sutanto and K Warwick, “Cluster Analysis for Multivariable Process Control,” Proceedings of American Control Conference, vol. 1, 1995, pp. 749-750

[3] Anil K Jain, M N Murty and P J Flynn, “Data Clustering: A Overview,” ACM Computing Surveys,

vol. 31, no. 3, 1999, pp. 265-323

[4] Anil K Jain, Robert P W Duin and Jianchang Mao, "Statistical Pattern Recognition: A Review," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, 2000, pp. 4-37

[5] Y M Sebzalli and X Z Wang, "Knowledge Discovery From Process Operational Data Using PCA and Fuzzy Clustering," Engineering Applications of Artificial Intelligence, vol. 14, 2001, pp. 607-616

[6] Timo Ahvenlampi and Urpo Kortela, "Clustering Algorithm in Process Monitoring and Control Application to Continuous Digester," Informatica, vol. 29, 2005, pp. 101-109

[7] Young-Hak Lee, Hyung Dae Jin, Chonghun Han, "On-Line Process State Classification for Adaptive Monitoring," Industrial Engineering Chemistry Research, 45, 2006, pp. 3095-3107

[8] 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

[9] 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, vol. 1, issue 2, 2007, pp. 108-118

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

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

[12] 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, no 9, issue 2, 2007, pp. 442-449

[13] D. T Pham et.al, "Data Clustering Using Bees Algorithm," Proceedings of 40th CIRP International Manufacturing Systems Seminar, 2007

[14] N Sujatha and K Iyakutty, "Refinement of Web Usage Data Clustering From K-Means with genetic Algorithm," European Journal of Scientific Research, vol. 42, no. 3, 2010, pp. 478-490

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

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

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

[18] V Kavitha and M Punithavalli, "Clustering Time Series Data Stream- A Literature Review," International Journal of Computer Science and Information Security, vol. 8, no. 1, 2010, pp. 289-294

[19] Izakian, H., Abraham, A., "Fuzzy C-Means and Fuzzy Swarm for Fuzzy Clustering Problem," Expert Systems with Applications, 2010, doi: 10.1016/j.eswa.2010.07.112

[20] S Kalyani and k S Swarup, "Supervised Fuzzy C Means Clustering Techniques for Security Assessment and Classification of Power System," International Journal of Engineering, Science and Technology, vol. 2, no. 3, 2010, pp. 175-185

[21] Xian-Xia Zhang et.al, "Spatially Constrained Fuzzy Clustering Based Sensor Placement for Spatiotemporal Fuzzy Control System," IEEE Transaction on Fuzzy Systems, vol. 18, no. 5, 2010, pp. 946-957

[22] Mika Liukkonen et.al, "Analysis of Flue Gas Emission Data From Fluidized Bed Combustion Using Self-Organizing Map," Applied Computational Intelligence and Soft Computing, Hindawi Publishing Corporation, 2010, pp. 1-8

[23] Vasil Simeonov et.al, "Lake Water Monitoring Data Assessment By Multivariate Statistics," Journal of Water Resource and Protection, vol. 2, 2010, pp. 353-361

[24] Ibrahim Massod and Adnan Hassan, "Issues in Development of ANN-Based Control Chart Pattern Recognition Schemes," European Journal of Scientific Research, vol. 39, no. 3, 2010, pp. 336-355