A New PCA Cluster-Based Granulated Algorithm Using Rough Set Theory for Process Monitoring

Hesam Komari Alaei[1], Seyed Iman Pishbin[1] and Karim Salahshoor
[1]*Research and development department, National Iranian Gas Company, Khorasan Razavi Province*
[2]*Department of Automation & Instrumentation Petroleum University of Technology, Tehran, Iran*
hesamkomari@yahoo.com, iman.pishbin@gmail.com, salahshoor@put.ir

Abstract

A new PCA algorithm is introduced, utilizing a rough cluster-based granulation scheme for segmentation of multivariate time series and process monitoring purposes. This granulated cluster-based algorithm can be used for segmentation of multivariate time series and initialization of other partitioning clustering methods that need to have good initialization parameters. The proposed algorithm is suitable for mining data sets, which are large both in dimension and size, in case generation. It utilizes Principal Component Analysis (PCA) specification and an innovative granular computing method for detection of changes in the hidden structure of multivariate time series data in a bottom up cluster merging manner. Rough set theory is used for feature extraction and solving superfluous attributes issue. The algorithm has been tested on an artificial case study. The resulting performances show the successful and promising capabilities of the proposed algorithm.

Keywords: *time series segmentation, Fault detection, Process monitoring, granular computing, Rough set, PCA, data mining, clustering*

1. Introduction

Most of partitioning clustering methods has the main problem of good initialization parameters, and without having this; the whole algorithm can diverge and become unstable or may converge to some local minimal points. To overcome the initialization problem, different methods for determining good initial parameters have been suggested, like subsampling, voting, two stage clustering and model-based clustering [3]. However, most of them have heavy computational requirements and/or sensitive to noise. Moreover most of these methods are working in space-domain and don't consider the time-domain characteristic, in another words in real process monitoring systems we usually encounter with multivariate time series data that should be segmented in time based on different products, faults and abnormal situations. Therefore considering time in our clustering base segmentation is an important topic which needs more attention. Another issue is that although segmentation of univariate time series data has been considered a lot, multivariate time series segmentation is rarely found [9,11]. In industrial processes, data-based approaches rather than model-based approaches have been widely used for process monitoring, because it is often difficult to develop detailed physical models. Different approaches for fault detection using mathematical models have been developed in the last 20 years. In this approach, the actual behavior of the process to be supervised is compared with that of a nominal model driven by the same inputs. Faults can be detected or isolated by evaluating the difference between the estimated value of

the model and the actual values of process variables. A general survey of supervision, fault detection and diagnosis methods in the model-based approaches is given by Isermann [1, 2].

Process plants now routinely have large volumes of historical data stored in databases. Although historical data consist of measurements on a large number of variables, these variables are highly correlated and the effective dimension of the space in which they move is very small. Typically only a few process disturbances or independent process changes routinely occur, and the hundreds of measurements on the process variables are only different reflections of these few underlying events. Multivariate statistical methods, Clustering base methods, Time series segmentation methods and Rough set theory are all different approaches that can be utilized in dealing with huge amount of historical data stored in industries. In recent years, chemometric techniques have been applied to monitoring and diagnosing in multivariate process, because a great number of variables are measured and they are highly correlated. Statistical process control (SPC) is well established for monitoring univariate processes, but they do not function well for multivariable processes. Jackson (1959) used principal component analysis (PCA) and proposed a T2 control chart for principal components. The workhorse of SPC control charts, such as Shewhart chart, CUSUM (cumulative sum) and EWMA (exponentially weighted moving average), applies well to monitoring process. Research efforts for more than a decade have focused on models developed by using latent variable methods such as PCA and projection to latent structures (PLS). Chemical engineers in academia and industry have applied PCA for abstracting structure from multidimensional chemical process data.PCA determines the most accurate lower dimensional representation of the data in terms of capturing the data directions that have the most variance. The resulting lower dimensional models have been used for detecting out-of-control status and for diagnosing disturbances leading to the abnormal process operation. Discriminant Partial Least Square (DPLS) computes a lower dimensional representation, which maximize the covariance between variables in that space and the predicted variables. Several researchers have applied DPLS and PCA to small-scale classification problems and compared their results. Latent variables exploit the main characteristic of process databases. The ability of PCA (Principal Component Analysis) to identify and capture process operating states through the measured variables and their interactions makes it a very powerful approach to statistical control [10].

In this paper PCA method is exploited to reduce historical process database in order to have better understanding of hidden information concealed in data. Then a new idea based on a granulation method is done on the PCA-transformed data. Since most of approaches were used for multivariate fault detection using the PCA method, are two statistics, T2 and Q. PCA granulated clustered-based algorithm that proposed in this article is applied for initialization of conventional T2 and Q methods that provide good comparing approaches. The results showed that novel proposed algorithm improved class separation and fault detection over conventional methods. Conventional PCA was used for data reduction; however improved PCA methods that introduced as Moving PCA, Multiscale PCA, can provide better performance. The proposed granulation method is introduced in section 4.This method encodes the available patterns in data base to obtain attribute value table. Then a reduced attribute value table is produced that has the number of occurrences of these repeated patterns. We consider these repeated patterns as reference points. Rough set theory was developed by Pawlak, for classificatory analysis of data tables. The main goal of rough set theoretic analysis is to synthesize approximation (upper and lower) of concepts from the acquired data. While fuzzy set theory assigns to each object a grade of belongingness to represent an imprecise set, the focus of rough set theory is on the ambiguity caused by limited discernibility of objects in the domain of discourse. A rough fuzzy granulation method has

also been introduced by Sankar K. Pal and Pabitra Mitra [5, 8]. A similarity matrix is introduced to give a measure for similarity of these reference points. This similarity matrix is then shrinking in an iterative manner using a bottom up algorithm and merges the most two similar reference points in each iteration. The paper is organized as follows. The PCA-based data transformation is described in Section 2. Rough set theory is explained in Section 3. The granulation method is introduced in Section 4. Similarity Matrix is explained in Section 5. Section 6 presents bottom-up cluster merging based on Similarity Matrix. Finally, the resulting PCA granulated clustered-based algorithm using rough set theory will be tested on an artificial case study introduced in [6].

2. PCA-Based Data Transformation

PCA is a useful statistical technique that has found application in different fields to find latent patterns in high dimensional data. A reduced representation of the original data is obtained which is smaller in size but having enough information to deal with. These variables are sorted upon their importance based on the magnitude of each eigenvalue, representing the variance in the direction of its corresponding eigenvector. So, the importance of each variable can be considered with its eigenvalues. PCA is defined as a linear transformation of the original correlated variables into a new set of variables that are uncorrelated with each other. The linear transformation is:

$$T = X.P \tag{1}$$

Where $X \in \Re^{n \times p}$, a matrix of n observations and p variables, measured about their means, the transformation matrix $P \in \Re^{p \times p}$ is called the loading matrix, which is the matrix of eigenvectors of the sample covariance or correlation matrix of the original data, and T is called score matrix, which is the projection of the original data into principal component subspace Utilizing PCA is beneficial in following aspects

- **(i)** A reduced representation of original data is obtained which is smaller in size but having enough information to deal with.
- **(ii)** Since PCA is a linear projection of data, the relative distances of data from each other remain the same.
- **(iii)** The transformed new variables that are principal Components (PCs) are uncorrelated.
- **(iv)** The correlation between observed variables can not be understood from them. Yet, PCA maps data points into new smaller dimensional space which is useful in the analysis and visualization of the correlated high dimensional data.
- **(v)** These variables are ordered regarding their importance and the magnitude of each eigenvalues describes the variance in the direction of its corresponding eigenvector. So the importance of each variable can be considered with its eigenvalues.

The starting point for PCA is the sample covariance matrix S (or the correlation matrix). For the p-variable problem, PCA method should be applied to normalized data. Therefore the first step in utilizing PCA method is typically 'auto scaling' or preprocessing the data to prevent dominance of relatively higher magnitude variables over other variables. Mean centering of the *n* variables in the columns and dividing them by the standard deviation has been done as follows [4]:

$$X = \frac{x_i - \mu_i}{\sigma_i} \tag{2}$$

Also a decision is made on the number of components to be retained. The retained components are significant contributors to the model. The greater the degree of correlation between the original variables, the fewer the number of PCs required. There are many ways of determining the number of PCs in batch-wise PCA, including: cross-validation [13], cumulative percent variance [14], average eigenvalues, imbedded error function and etc. In this paper, cumulative percent variance (CPV) method is used to determine the number of PCs. The CPV is a measure of the percent variance captured by the first l_k PCs:

$$CPV(l_k) = \frac{\sum_{j=1}^{l_k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} 100\% \tag{3}$$

The number of PCs is hence chosen when the CPV reaches a predetermined limit, say 95%.

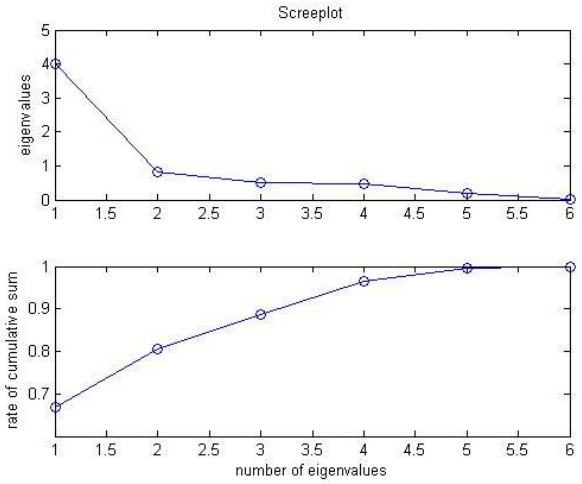

Figure 1. Scree-plots of the Artificial Data Sets in [6]

3. Rough Set Theory

An information granule is a clump of objects in the universe of discourse drawn together by indistinguishability, similarity, proximity, or functionality. Information granules reduce the data by identifying equivalence classes, i.e. objects that are indiscernible, using the available attributes. Only one of the elements of the equivalence class is needed to represent the entire class. Hence, rough set theory is a natural choice for case selection in domains which are data rich, contain uncertainties, and allow tolerance for imprecision [12]. Attribute value table is equipped with decision column that, constructing based on normal and abnormal conditions. The equivalence classes are indicated as granules. Rough fuzzy granulation with Gaussian membership function provides the ability to specify lower and upper approximations based on average and standard deviation of Gaussian distribution. This ability is used to facilitate clarity of ambiguity with high accuracy. Table 1 indicates that

similar data (features) with the same decision rule as normal or abnormal, can delineate concept crisply, i.e., delineate the objects which definitely "belong" to the concept of normal or abnormal condition or "do not belong". Despite of that, similar data with opposite decision rule, indicates that the concept cannot delineate crisply. So, these granules belong to both normal and abnormal condition. These granules are called ambiguity. To clarify ambiguity, the proposed cluster based granulated algorithm is used.

Table 1. Attribute Value Table

Feature 1	…	**Feature n**	**Decision**
Sensor1(t1)	…	Sensorn(t1)	Normal
…	…	…	…
Sensor1(tN)	…	Sensorn(tN)	Abnormal

4. The Granulation Method

In this section, a method of obtaining the granular feature space is described. Consider a database in a form $X_{N\times n}$, that N is the number of patterns (samples) and n is the number of attributes (features) in PCA space which are actually our eigenvectors. Let a pattern (object) be represented as

$$F = [\, f_1, f_2, \ldots, f_n \,] \tag{4}$$

Let's assign some definite points as representative of all other points on each feature (variable). We define these representative points as the center of clusters for each feature. It means that each point is assigned to the nearest cluster center on that feature so that we have to deal with only these centering points.

Kmeans algorithm is used for clustering of each feature. It is obvious that for this purpose the number of clusters should be defined. Different approaches are suggested in literature [3]. A silhouette criterion is used for this purpose, which has been explained in [3] as shown in Figure 2.

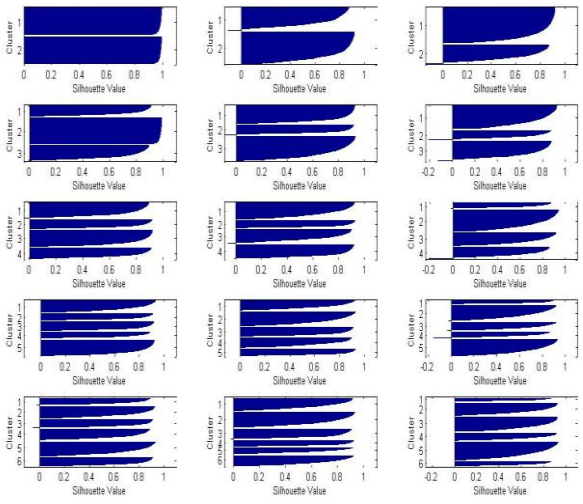

Figure 2. A silhouette plot of the first 3 PCs of case study for different cluster centers (2, 3, 4, 5, 6 clusters).

Let a pattern in this new data base represented as:

$$\hat{F} = [\, \hat{f}_1^{1,\ldots,K_1} , \hat{f}_2^{1,\ldots,K_2} , \ldots , \hat{f}_n^{1,\ldots,K_N} \,] \tag{5}$$

Which K_1, K_2, …, K_n are the number of clusters for each feature (PCs) . Now the entire data base is mapped into some defined points. Actually a granulation method has been applied to the PCA database. We call the obtained new data base *'Attribute Value Table'.* This table may have some repeated patterns. Therefore, a reduced attribute value table is produced that has the number of occurrences of these repeated patterns. We consider these repeated patterns as reference points.

5. Similarity Matrix

A similarity matrix that contains similarity of each of reference pattern to other reference points has been defined. Similarity matrix is an upper triangular matrix that any element of it contains the similarity of corresponding row and column. Therefore, the similarity vector of each point is defined as the corresponding row vector in a similarity matrix.

For defining the similarity criterion between two patterns there are some consideration. Features of each pattern correspond to principal components, it means, first feature which correspond to first PCs can be considered to have importance of its eigenvalues (λ) of PCA model and the same for other features (PCs) which have been used, therefore this fact is noticed for defining the similarity of patterns.

$$S_{i,j} = \frac{(\hat{F}_i - \hat{F}_j)\,\lambda_{norm}}{\max(\hat{F}) - \min(\hat{F})} \tag{6}$$

Note that this formula is in compact form and all terms are vectors. where

$$\hat{F}_i = [\, \hat{f}_{i1} \quad \hat{f}_{i2} \quad \ldots \quad \hat{f}_{in} \,],\ \hat{F}_j = [\, \hat{f}_{j1} \quad \hat{f}_{j2} \quad \ldots \quad \hat{f}_{jn} \,] \tag{7}$$

$$\lambda_{norm} = \frac{[\,\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_n\,]}{\sum_{i=1}^{n} \lambda_i} = [\,\lambda_1' \quad \lambda_2' \quad \ldots \quad \lambda_n'\,] \tag{8}$$

Therefore the expanded formula:

$$S_{i,j} = Sum[\ \frac{(\hat{f}_{i1} - \hat{f}_{j1})\,\lambda_1'}{\max\limits_{x=1}^{k_1}\{\hat{f}_{x_1}\} - \min\limits_{x=1}^{k_1}\{\hat{f}_{x_1}\}} \qquad \frac{(\hat{f}_{i2} - \hat{f}_{j2})\,\lambda_2'}{\max\limits_{x=2}^{k_2}\{\hat{f}_{x_2}\} - \min\limits_{x=1}^{k_2}\{\hat{f}_{x_2}\}} \quad \ldots] \tag{9}$$

As shown for measuring the similarity of two patterns we subtract two patterns from each other. To consider the importance of features, vector of normal eigenvalues is multiplied. The denominator term which contains difference from maximum to minimum cluster centers for each feature is considered for normalization. In this criterion not only features of same clusters has cost function of zero but also for different cluster centers the amount of difference of one cluster centers from the other one is take into account.

6. Bottom-Up Cluster Merging based on Similarity Matrix

In data mining, the Bottom-Up segmentation algorithm has been extensively used to support a variety of data mining tasks, [7]. Keogh et al. compare this method of clustering with two other general methods of clustering time series data, means Top-Down and Sliding Window methods, and conclude that Bottom-Up segmentation work better than the other two methods almost on all data bases [7].

The idea of Bottom-Up segmentation is used here. The similarity matrix is shrinking by using a Bottom-Up manner. Let's find the minimum similarity amount in this matrix, means the most similar reference points among available points. The algorithm merges these two reference points (clusters) by substituting their similarity vector with a new similarity vector which represents the similarity of this new cluster with other remaining points. The new similarity vector is defined as follows:

$$S^{*}_{i,j} = \frac{N_i S_i + N_j S_j}{N_i + N_j} \tag{10}$$

Where N_i, N_j are number of occurrence of i, j reference points and S_i, S_j are similarity vectors shows the similarity of points i, j to other reference points. The iterative cluster merging algorithm continues until the minimum similarity cost of merging reference points is less than a threshold. This threshold is usually between 0.3 and 0.6 depending on homogeneity of the time series.

During hierarchical bottom-up algorithm the number of reference points is reduced and also each point in time will be dedicated to one of these reference points. Note that number of reference points means number of different regions in PCA space that can be identical to different products or maybe faults or abnormal situation.

Diagram of the proposed algorithm is illustrated in figure 3. The figure shows that the proposed algorithm utilizes the properties of PCA for revealing the hidden information concealed in data. Rough set theory reduces the data by identifying equivalence classes, i.e. objects that are indiscernible, using the available attributes. The equivalence classes are indicated as granules. The similar data (features) with the same decision rule as normal or abnormal can delineate concept crisply and similar data with opposite decision rule, indicates that the concept cannot delineate crisply. These granules are called ambiguity. To clarify ambiguity, the cluster based granulated algorithm is used. Attribute value table is constructed based granulation algorithm by considering the normal, abnormal an ambiguity features obtained by rough set theory. Attribute value table has the number of occurrences of repeated patterns, that these repeated patterns are considered as reference points. A similarity matrix that contains similarity of each of reference pattern to other reference points has been defined. Finally, bottom-up cluster merging based on similarity matrix is implemented.

The combination of PCA with rough granulation method has led to powerful algorithm that could detect hidden changes in multivariate time series. The implementation of the proposed algorithm on an artificial data base shows promising results.

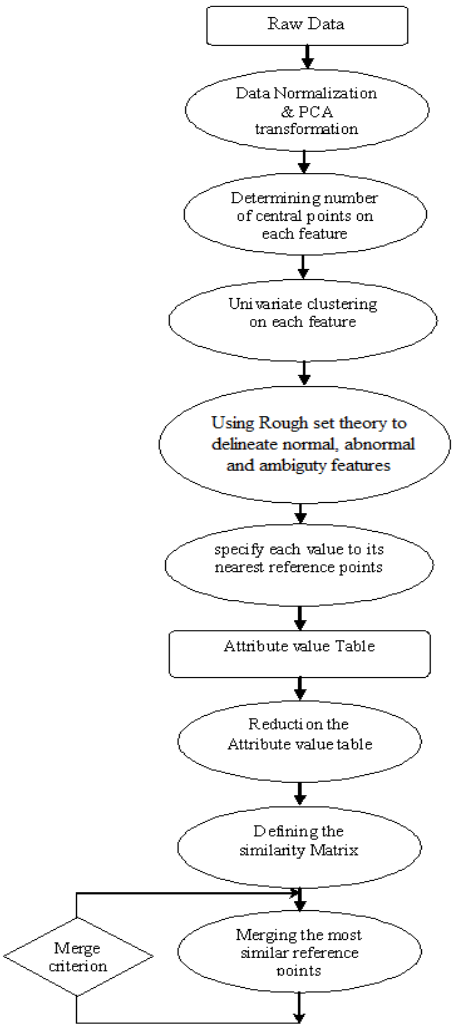

Figure3. Diagrams of Proposed Algorithm

7. Case Study

In our work the case study which has introduced by Janos Abonyi et al. [6] has been used. It is in fact an artificial dataset. Consider the synthetic dataset shown in Figure5. The 2000 observed variables that can be seen on Figure5 are not independent; they were generated by the latent variables shown in Figure 4 The correlation among the observed variables changes at the quarter of the time period while the mean of the latent variables changes at the half of the time period. These changes are marked by the vertical lines in Figure 4.

As can be seen on Figure5, such information can be detected neither by application of univariate segmentation algorithms, nor by the visual inspection of the observed variables. Hence developing an algorithm that is able to handle time varying characteristics of multivariate data is essential to detect:
(i) Changes in the mean; (ii) changes in the variance; and (iii) changes in the correlation structure among the multivariate data [6]. To discover these types of changes, being hidden

inherently in the relationships of multivariate time series, multivariate statistical tools should be applied by the segmentation algorithm.

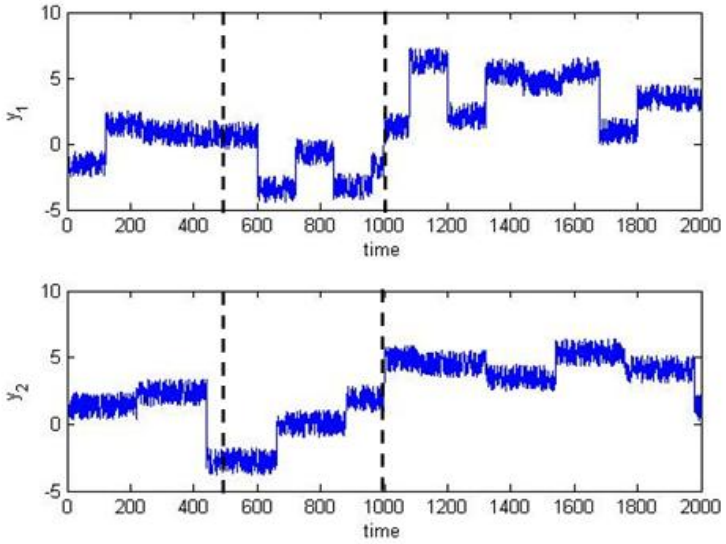

Figure 4. Latent Variables Data

In this case study, cumulative percentage for the significant two PCs is 96.381 and for the three PCs are 98.899. Therefore, the original (2000*6) data-base can be projected to 2000*3 in the PCA space.

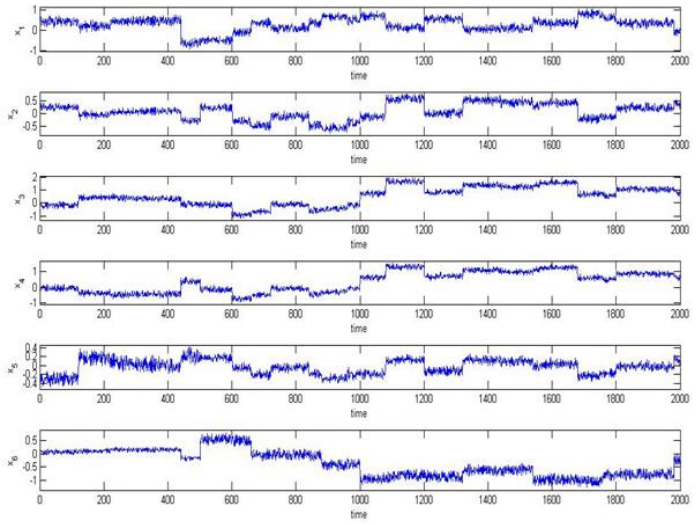

Figure 5. Observed Variable

Silhouette criterion is used for determining the desired number of clusters for each PC. So, Kmeans classifier is applied to construct an attribute value table, being followed by a merging algorithm afterwards. The result of the proposed novel algorithm is shown in Figure 6. This figure shows the first time series segmentation which suggests 16 granules.

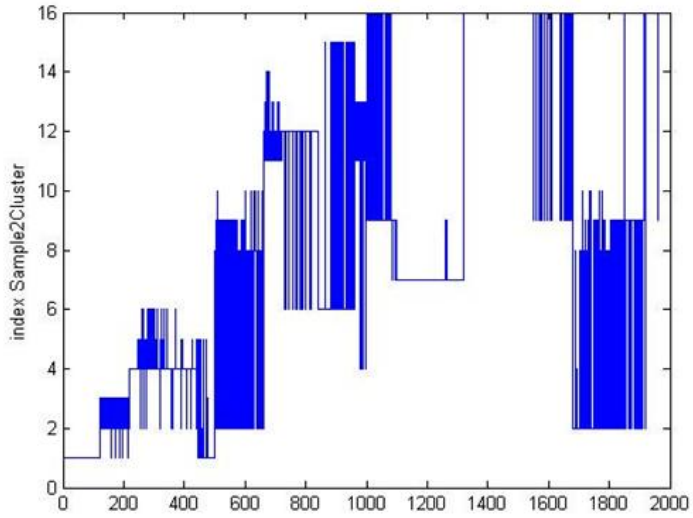

Figure 6. The first time series segmentation suggest 16 granules (reference points) that each point is dedicated to one of these reference points.

Figure 7. shows that the cluster-based granulation algorithm can monitor the time series well. As it has been demonstrated in Figure 7, both the changes made at the quarter and the half of the test time period can accurately be detected.

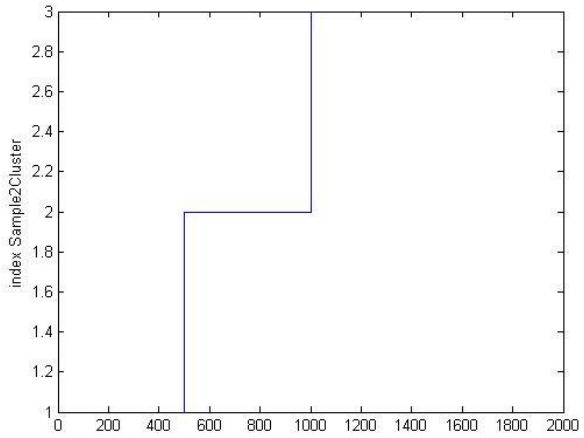

Figure 7. The Segmentation Algorithm after 13 Times of Running Merging Algorithm

8. Conclusion

This paper presented a new Granulated Cluster-Based Algorithm using rough set theory for segmentation of multivariate time series and monitoring of industrial process. It utilizes the properties of PCA for revealing hidden information concealed in data. The combination of PCA with granulation methods enhanced with rough set theory leads to the algorithm that could detect hidden changes in multivariate time series. The implementation of algorithm on artificial data base shows promising results that could detect the change in mean and correlation of original data perfectly.

References

[1] Rolf Isermann, MODEL-BASED FAULT DETECTION AND DIAGNOSIS –STATUS AND APPLICATION-

[2] Rolf Isermann, 2005. FAULT DIAGNOSIS SYSTEMS An introduction from fault detection to fault tolerance.

[3] Wendy L. Martinez, Angel R. Martinez, Exploratory Data Analysis with Matlab, Computer Science and Data Analysis Series

[4] Y.M. Sebzalli, X.Z. Wang, Knowledge discovery from process operational data using PCA and fuzzy clustering, Engineering Applications of Artificial Intelligence 14 (2001) 607–616

[5] Sankar K. Pal, and Pabitra Mitra, Case Generation Using Rough Sets with Fuzzy Representation, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 3, MARCH 2004

[6] Janos Abonyi, Balazs Feil, Sandor Nemeth, Peter Arva, Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series ,Fuzzy Sets and Systems 149 (2005) 39–56

[7] Eamonn Keogh, Selina Chu, David Hart, Michael Pazzani, An Online Algorithm for Segmenting Time Series.IEEE Intenat.Conf. on Data Mining,

[8] Pabitra Mitra , Sankar K. Pal , Md Aleemuddin Siddiqi, Non-convex clustering using expectation maximization algorithm with rough set initialization Pattern Recognition Letters 24 (2003) 863–873

[9] Kaufman, Leonard and Peter J.Rousseeuw.1990. Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley & Sons

[10] Theodora kourti, Process analysis and abnormal situation detection: From theory to practice. IEEE Control Systems Magazine, October 2002

[11] Janos Abonyi, Balazs Feil, Sandor Nemeth, Peter Arva, Principal Component Analysis based Time Series Segmentation {A New Sensor Fusion Algorithm.

[12] Skowron and C. Rauszer, "The Discernibility Matrices and Functions in Information Systems," Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, R. Slowi_nski, ed. pp. 331-362, Dordrecht: Kluwer Academic, 1992.

[13] U.Kaymak, R.Babuska, compatible cluster merging for fuzzy modeling,in: Proc, IEEE internat. Conf. on fuzzy systems, Yokohama, Japan, 1995, 897-904

[14] E. R. Malinowski. "Factor Analysis in Chemistry". Wiley-Interscience: New York (1991).

Authors

Hesam Odin Komari Alaei received the BS degree in Electrical Communication Engineering from the Sadjad University of Mashhad, Iran. He obtained the MS Degree in Automation and Instrumentation from the Petroleum University of Technology. His research interests include data mining and sensor fusion, pattern recognition, statistical process control, computer vision, intelligent control, fault tolerant control, fault monitoring and control and advance process control.

Seyed Iman Pishbin is the head of Research and Development Department, National Iranian Gas Company, Khorasan Razavi Province. He received the B.S. Degree from Tarbiat Modares University, Iran, and the M.S. and Ph.D. Degrees in Mechanical Engineering - Energy Conversion Field from the Ferdowsi University of Mashhad, Iran.

Karim Salahshoor is an Associate Professor of the Automation and Instrumentation Department, Petroleum University of Technology, Tehran, Iran. He received the B.S. Degree from AIT, Abadan, Iran, and the M.S. and Ph.D. Degrees from the UMIST, Manchester, England. His current research interests include industrial networking for process instrumentation and control systems, fieldbus, and advanced process control applications.