International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-4, April 2012

Data Driven Multivariate Technique for Fault Detection of Waste Water Treatment Plant

Subhransu Padhee, Nitesh Gupta, Gagandeep Kaur

Abstract - Collection of raw data from different sensors, processing the data and extracting information from it is a very challenging task. Because of the enhanced memory capacity of the present day computers, data logging has reached to a new level. The analyst has to classify the data according to their traits from the offline logged data. The whole task of collection of raw data, classification of data according to their traits involves different statistical as well as soft computational techniques. This research paper takes a case study of waste water treatment plant and using different data driven multivariate statistical techniques and soft computational techniques determine the faults in the system. This paper uses principal component analysis and backpropagation algorithm to classify the data and detect the faults in a waste water treatment plant.

Index Terms - backpropagation, multivariate statistical technique, principal component analysis.

I. INTRODUCTION

In a process control application, identification of process structure and parameters, monitoring the quality of the process outputs and detection of abnormal trends in different variables, is an upcoming research area. Collection of raw data from different sensors, processing the data and extracting valuable information about the condition of the process is a very challenging task. Because of the enhanced memory capacity of the present day computers, data logging has reached to a new level. The analyst has to classify the data according to their traits from the offline logged data. The analyst needs to understand the process dynamics with the help of historical data or logged data. Different statistical and machine learning algorithms are used to determine the faulty trends in process dynamics. To operate the plant in an efficient manner and to obtain quality product in economical way, fault detection in process dynamics is important. The whole new branch of statistical process control came in to existence, because in modern day monitoring and fault diagnosis of process control application are in huge demand because every industry needs to give consistent output with the best of the quality. Looking in to the complexity of the process industry, multivariate statistical process control is used for the above purpose.

The objective of the study is to use statistical and soft computational based technique to model the process and monitor of the quality of process variables. The main objectives are

- Identification of process dynamics
- Quality monitoring
- Fault detection

Identification of process dynamics can be obtained from system identification tools. Different multivariate statistical techniques like PCA, ICA, PLS can be used for dimensionality reduction and backpropagation algorithms and variants of BPA, SVM can be used as an classifier which classifies different kind of data.

The content of the paper is divided in to 6 sections. Section II describes the fault detection system. Section III describes the technique of dimensionality reduction and PCA. Section IV investigates the previous work done in multivariate statistical analysis to detect faults in process control application. Section V takes a case study of waste water treatment plant with large number of inputs and output variables and finds the fault in the system by statistical and soft computational techniques. Section VI concludes the paper.

II. FAULT DETECTION SYSTEM

Fault detection in a process plant can be performed using the following techniques.

1. Data driven technique
2. Analytical technique
3. Knowledge based technique

2.1 Data driven technique

Data driven technique uses online or offline data to detect the faults and the output mainly depends upon the quality and quantity of data. This technique uses univariate and multi variate statistical techniques for dimensionality reduction and soft computing based classification technique (supervised and unsupervised) technique to classify between healthy and faulty data. The data driven technique uses statistical techniques like Principal Component Analysis (PCA) to reduce the high dimensional data obtained from the computer based manufacturing system to lower dimension data. The main disadvantage of the technique is that it is dependent on the quality and quantity of data and might give wrong results if there is abnormality in data or missing values.

The data driven techniques can

Manuscript published on 30 April 2012.
* Correspondence Author (s)
Subhransu Padhee, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala, India, (e-mail: subhransu_padhee@yahoo.com).
Nitesh Gupta, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala, India, (e-mail: niteshgupta531@gmail.com).
Gagandeep Kaur, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala, India, (e-mail: gagan_deep_kaur@yahoo.com).

be broadly classified as

1. Univariate technique
2. Multivariate technique

2.1.1 Univariate technique

In univariate technique, each of the variables of the process plant is analysed independently. It is an effective way of analysing each and every variable of the dataset and their response with respect to time. But if the said variable which is under consideration of univariate technique is related to any other variable of the process by some means or the other, then the univariate analysis fails.

2.1.2 Multivariate technique

When the univariate technique fails, multi variate technique comes in to existence. Using multivariate technique, a large number of variables of the process can be analyzed. The multivariate technique gives the list of variables which are the primary variables of the process and the other variables of the process are the mere reflection of the primary variable of the process. Out of hundreds of process parameters, the researcher can reduce the number of variables to two-three using multivariate technique. These couple of variable represents the most of the percentage of data.

There are different multivariate analysis techniques, such as

1. Factor analysis
2. Cluster analysis
3. Multidimensional scaling
4. T^2 statistics
5. Principal component analysis
6. Independent component analysis

2.2 Analytical technique

Analytical approach uses mathematical model to detect the faults. This technique is good for small scale system but very expensive for large scale system.

2.3 Knowledge based technique

Knowledge based techniques uses qualitative methods for fault detection. This approach is suitable for system without detailed mathematical models.

III. Dimensionality Reduction and PCA

In modern day computer based process control, there are hundreds and thousands of process variables. These variables are interrelated to each other and each of the variables contributes to the quality of output product.

So, it is very difficult to comprehend the associated patterns in the dataset. Hundreds of the variables make the system redundant, so it also makes the classification task more challenging. This results in high level of correlation and multi collinearlty.

There are different techniques to reduce the dimensionality of the dataset such as

1. Multiple regression
2. Discrimination analysis
3. Principal component analysis

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique to reduce the dimension of the dataset, but it preserves the correlation structure between the variables. By application of PCA one can preserve most of the variation of the process data.

When PCA is applied in a very large dataset, the data will be transformed in such a way that the 90% of the variance of the replicated in the first few principal components only.

PCA can be calculated in different ways

1. Correlation Method
2. Covariance Method

PCA is an orthogonal linear transformation that transforms the data in to a new coordinate system such that the greatest variance by any projection of data comes to lie in the first coordinate, called first principal component, the second greatest variance in the second coordinate called second principal component and so on.

3.2 Calculation of PCA

PCA can be calculated using the following steps

1. Input data
2. Subtract the mean
3. Calculate covariance matrix
4. Calculate Eigen values and Eigen vectors of the covariance method
5. Choosing components and forming a feature vector
6. Deriving a new dataset

IV. Related Works

There is plenty of literature where researchers implement supervised as well as unsupervised based classification techniques for classification and detection of faults.

4.1 Multivariate Techniques

R Dunia et.al, has proposed the use of PCA for sensor fault identification. The principal component model captures measurement correlation and reconstructs the variables using optimization techniques. The status of the sensor is determined by sensor validity index [1].

H Albazzaz et.al, performed a comparative study of multidimensional visualization techniques and multivariate statistical process control for process historical data analysis. The study was carried out for a waste water treatment plant. [9].

V Venkatasubramaniam et.al in his pioneer work reviews different quantitative model based methods [5], qualitative methods and search strategies [6] and process history based methods [7] to efficiently detect process faults.

A AlGhazzawi et.al, has implemented multivariate statistical process control to have a online fault detection system in a petroleum refining process [14]

D Wang et.al has proposed that Robust Multi Scale PCA improves the performance of MSPCA technique and uses RMSPCA models for process monitoring purposes [11].

4.2 Unsupervised classification (clustering)

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

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

three hundred data from the process site and applying principal component analysis and fuzzy c means clustering algorithm in the datasets [3].

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 [15].

V. CASE STUDY

This research paper considers a waste water treatment plant and ascertains the condition of the plant (Healthy or Faulty) by using multivariate statistical techniques and artificial neural network.

The dataset for waste water treatment plant is taken from UCI machine learning respiratory; and is available online **http://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant** (last accessed 4th March 2012). This is a multivariate dataset created by Manel Poah, Unitat d'Enginyeria Quimica Universitat Autonoma de Barcelona. Bellaterra. Barcelona; Spain and donated by Javier Bejar and Ulises Cortes, Dept. Llenguatges i Sistemes Informatics; Universitat olitecnica de Catalunya. Barcelona; Spain. The dataset shows the recorded reading of the variables for 527 days.

This is an activated sludge process located and treats a daily flow of 35000 m^3 of domestic and industrial wastewater. This plant consists of 3 parts

1. Pre treatment
2. Primary Treatment
3. Secondary treatment

Table 1: Different attributes of waste water treatment plant

Attributes	Description
Q-E	Input flow to plant
ZN-E	Input zinc to plant
PH-E	Input ph to plant
DBO-E	Input biological Demand of O_2 to Plant
DQO-E	input chemical demand of oxygen to plant
SS-E	input suspended solids to plant
SSV-E	input volatile supended solids to plant
SED-E	input sediments to plant
COND-E	input conductivity to plant
PH-P	input pH to primary settler
DBO-P	input Biological demand of oxygen to primary settler
SS-P	input suspended solids to primary settler
SSV-P	input volatile supended solids to primary settler
SED-P	input sediments to primary settler
COND-P	input conductivity to primary settler
PH-D	input pH to secondary settler
DBO-D	input Biological demand of oxygen to secondary settler
DQO-D	input chemical demand of oxygen to secondary settler
SS-D	input suspended solids to secondary settler
SSV-D	input volatile suspended solids to secondary settler
SED-D	input sediments to secondary settler
COND-D	input conductivity to secondary settler
PH-S	output pH
DBO-S	output Biological demand of oxygen
DQO-S	output chemical demand of oxygen
SS-S	output suspended solids
SSV-S	output volatile suspended solids
SED-S	output sediments
COND-S	output conductivity
RD-DBO-P	performance input Biological demand of oxygen in primary settler
RD-SS-P	performance input suspended solids to primary settler
RD-SED-P	performance input sediments to primary settler
RD-DBO-S	performance input Biological demand of oxygen to secondary settler
RD-DQO-S	performance input chemical demand of oxygen to secondary settler
RD-DBO-G	global performance input Biological demand of O_2
RD-DQO-G	global performance input chemical demand of O_2
RD-SS-G	global performance input suspended solids
RD-SED-G	global performance input sediments

Table 1 indicates different inputs and output variables of the waste water system. The maximum, minimum and mean value of each variable is illustrated in table 2.

Table 2: Maximum, Minimum and Mean values of variables

Variables	Max	Min	Mean
Q-E	10000	60081	37226.56
ZN-E	0.1	33.5	2.36
PH-E	6.9	8.7	7.81
DBO-E	31	438	188.71
DQO-E	81	941	406.89
SS-E	98	2008	227.44
SSV-E	13.2	85	61.39
SED-E	0.4	36	4.59
COND-E	651	3230	1478.62
PH-P	7.3	8.5	7.83
DBO-P	32	517	206.2
SS-P	104	1692	253.95
SSV-P	7.1	93.5	60.37
SED-P	1	46	5.03
COND-P	646	3170	1496.03
PH-D	7.1	8.4	7.81
DBO-D	26	285	122.34
DQO-D	80	511	274.04
SS-D	49	244	94.22
SSV-D	20.2	100	72.96
SED-D	0	3.5	0.41
COND-D	85	3690	1490.56
PH-S	7	9.7	7.7
DBO-S	3	320	19.98
DQO-S	9	350	87.29
SS-S	6	238	22.23
SSV-S	29.2	100	80.15
SED-S	0	3.5	0.03
COND-S	683	3950	1494.81
RD-DBO-P	0.6	79.1	39.08
RD-SS-P	5.3	96.1	58.51
RD-SED-P	7.7	100	90.55
RD-DBO-S	8.2	94.7	83.44
RD-DQO-S	1.4	96.8	67.67
RD-DBO-G	19.6	97	89.01
RD-DQO-G	19.2	98.1	77.85
RD-SS-G	10.3	99.4	88.96
RD-SED-G	36.4	100	99.08

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The dataset provided in UCI machine learning respiratory has some missing data. The number of missing data is illustrated in table 3.

Table 3: Number of missing variables in each parameter of the dataset

Variables	Number of Missing Variables
Q-E	18
ZN-E	3
PH-E	0
DBO-E	23
DQO-E	6
SS-E	1
SSV-E	11
SED-E	25
COND-E	0
PH-P	0
DBO-P	40
SS-P	0
SSV-P	11
SED-P	24
COND-P	0
PH-D	0
DBO-D	28
DQO-D	9
SS-D	2
SSV-D	13
SED-D	25
COND-D	0
PH-S	1
DBO-S	23
DQO-S	18
SS-S	5
SSV-S	17
SED-S	28
COND-S	1
RD-DBO-P	62
RD-SS-P	4
RD-SED-P	27
RD-DBO-S	40
RD-DQO-S	26
RD-DBO-G	36
RD-DQO-G	25
RD-SS-G	8
RD-SED-G	31

The missing values are interpolated with the help of bicubic interpolation technique. It is very hard to classify a large dataset, so dimensionality of the dataset is reduced. To achieve dimensionality reduction, statistical tool like PCA is used. After dimensionality reduction of the dataset two principal component of the dataset is found out. After dimensionality reduction, neural network based classification technique is used to differentiate between healthy and faulty data. In neural network based classification technique, backpropagation algorithm is used. The flow chart of operation is shown in figure 2. Figure 1 shows the box plot diagram of the dataset, which graphically represents the data measure in an interval scale. It is a type of graph which is used to show the shape of the distribution, its central value, and variability and it is used in exploratory data analysis.

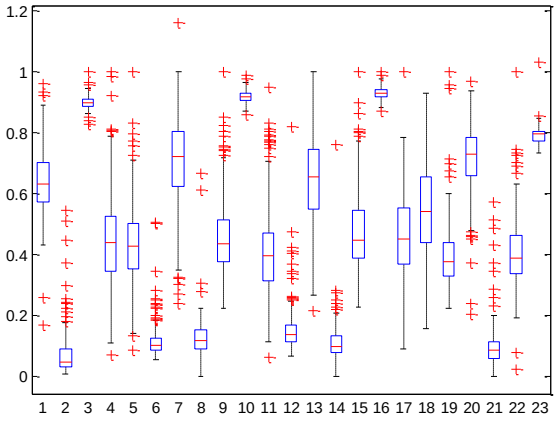

Figure 1: Box plot representation of dataset.

Figure 2 represents the flow chart to classify the dataset in to two categories, healthy or faulty. The first step is pre processing of data, then dimensionality reduction of data and then neural network based classification to classify the dataset.

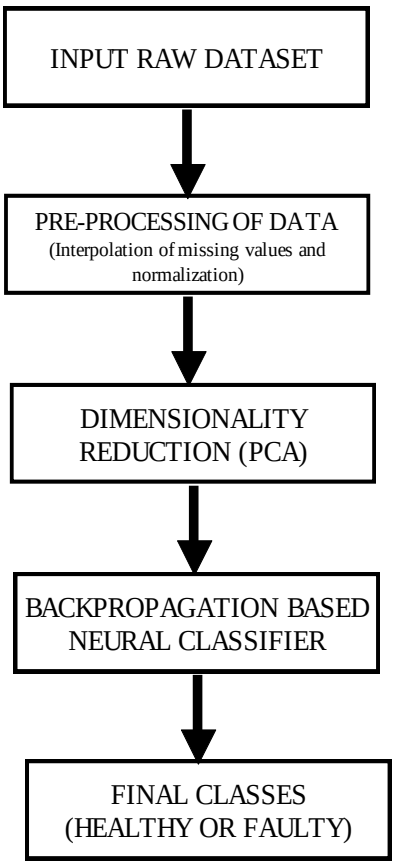

Figure 2: Block diagram for classification

After pre-processing of data, PCA is implemented in the dataset to reduce the dimension. PCA gives two variables, PC1 and PC2 which imitates 98.6% and 1.4% of the dataset. Figure 3 shows the scatter plot of PC1 and PC2.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

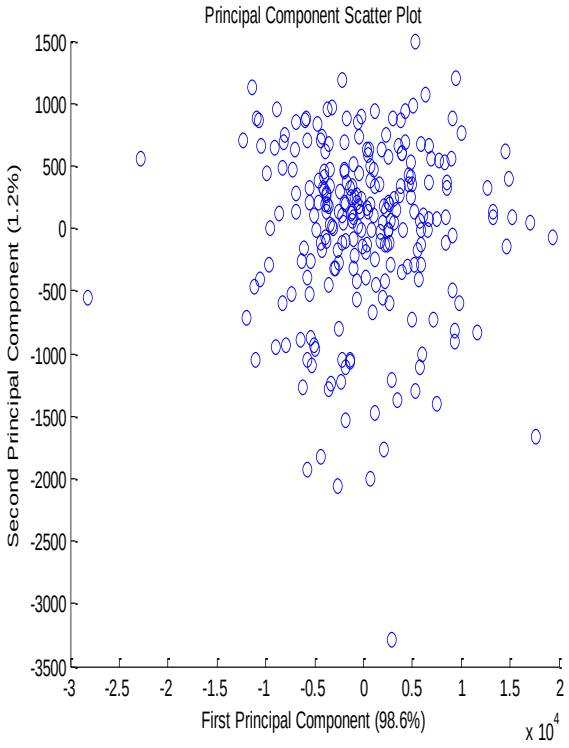

Figure 3: Scatter plot for PC1 and PC2

In PC1, 98.6% of variance of the dataset is represented where as in PC2, 1.4% of the variance of dataset is represented. By the help of PCA, very large dataset is reduced to only two variables which replicate 100% variance of the dataset.

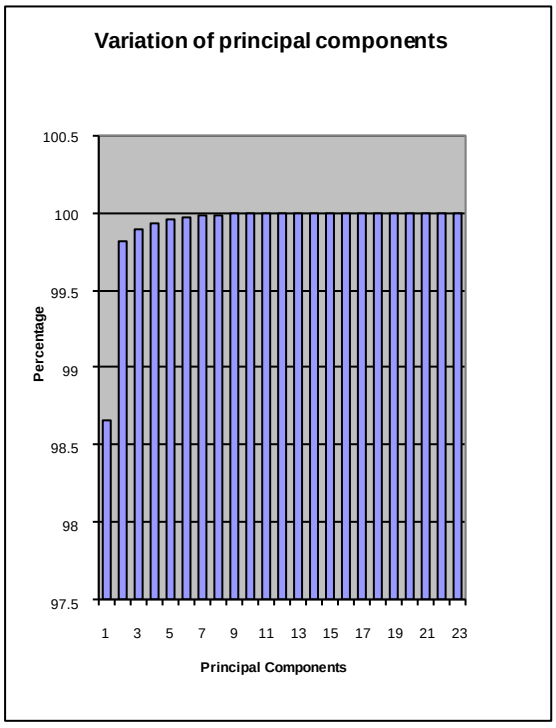

Figure 4: Different principal components and the variation of principal components.

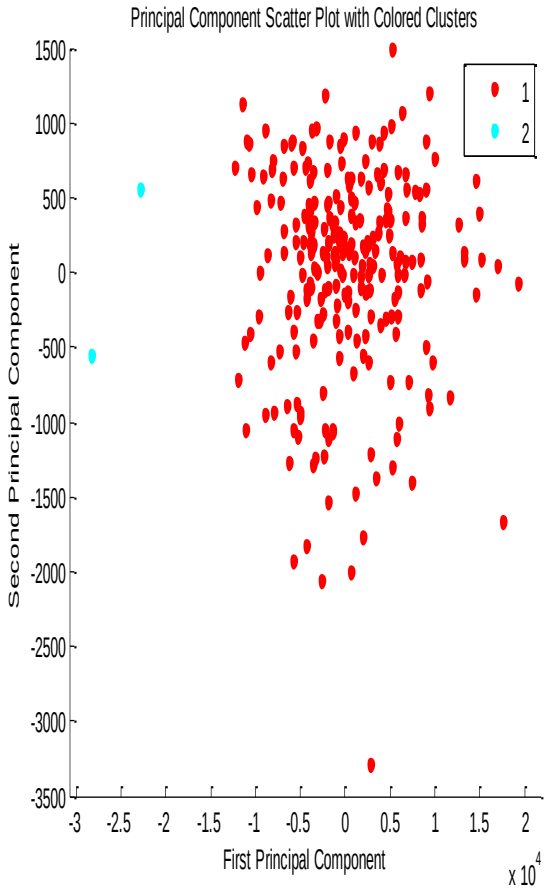

Figure 5: Classification of variables using PCA

Figure 4 shows the amount of variance of dataset held by each principal component. Figure 5 represents the classification of data of the dataset using PCA

VI. CONCLUSION

This paper proposes an efficient approach to detect different faults in process dynamics. The data driven approach uses statistical multivariate approach like PCA and soft computational algorithm like feed forward architecture with back propagation algorithm to classify healthy and faulty data. The process under consideration is a waste water plant with a large number of inputs and output variables. The dataset records the data for 527 numbers of days and is available at UCI machine learning respiratory. First the data is pre processed, then dimensionality of the dataset is reduced and at the end backpropagation algorithm is used to classify the data.

There is a lot of future scope of the study. Different multivariate statistical techniques like ICA and hybrid statistical multivariate techniques like PCA+ICA can be used for dimensionality reduction. Different algorithms in backpropagation can also be used and can be compared.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

REFERENCES

[1] R Dunia, S J Qin, T F Edgar and T J McAvoy, "Identification of faulty sensors using principal component analysis," AICHE Journal, vol. 42, no. 10, 1996, pp. 2797- 2812

[2] T Amand, G Heyen and B Kalitventzeff, "Plant monitoring and fault detection: Synergy between data reconciliation and principal component analysis," Computers and Chemical Engineering, vol. 25, no. 4-6, 2001, pp. 501-507

[3] 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

[4] N Bendwell, "Monitoring of a waste water treatment plant with a multi variate model- The benefits of PCA technology explained," Pulp and Paper-Canada, vol. 103, no. 7, 2002, pp. 43-46

[5] V Venkatasubramaniam, R Rengaswamy, K Yin and S N Kavuri, "A review of process fault detection and diagnosis Part-I: Qulitative model based methods," Computers and Chemical Engineering, vol. 27, no. 3, 2003, pp. 293 – 311

[6] V Venkatasubramaniam, R Rengaswamy and S N Kavuri, "A review of process fault detection and diagnosis Part-II: Quantitative model and search strategies," Computers and Chemical Engineering, vol. 27, no. 3, 2003, pp. 313 – 326

[7] V Venkatasubramaniam, R Rengaswamy, S N Kavuri and K Yin, "A review of process fault detection and diagnosis Part-III: Process history based methods," Computers and Chemical Engineering, vol. 27, no. 3, 2003, pp. 327 – 346

[8] R Ganesan, T K Das, V Venkataraman, "Wavelet based multiscale statistical process monitoring: A literature review," Iie Transactions, vol. 36, no. 9, 2004, pp. 787 – 806

[9] H Albazzaz, X Z Wang and F Marhoon, "Multidimensional visualization for process historical data analysis: a comparative study with multivariate statistical process control," Journal of Process Control, vol. 15, no. 3, 2005, pp. 285-294

[10] X Sun, H J Marquez, M Riaz and T W Chen, "An improved PCA method with application to boiler leak detection," ISA Transactions, vol. 44, no. 3, 2005, pp. 379 - 397

[11] D Wang and J A Romagnoli, "Robust multi scale principal components analysis with applications to process monitoring," Journal of process control, vol. 15, no. 8, 2005, pp. 869 – 882

[12] H Albazzaz, X Z Wang, "Introduction of dynamics to an approach for batch process monitoring using independent component analysis," Chemical Engineering Communications, vol. 194, no. 2, 2007, pp. 218 – 233

[13] Z Q Ge and Z H Song, "Process monitoring based on independent component analysis-principal component analysis and similarity factors", Industrial & Engineering Chemistry Research, vol. 46, no. 7, 2007, pp. 2054 - 2063

[14] A. AlGhazzawi and B. Lennox, "Monitoring a complex refining process using multivariate statistics," Control Engineering Practice, vol. 16, no. 3, 2008, pp. 294-307

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

[16] 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.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)