Malfunction Diagnosis in Industrial Process Systems using Data Mining for Knowledge Discovery

E. Lithoxoidou¹, C. Ziogou², T. Vafeiadis¹, S. Krinidis¹, D. Ioannidis¹, S. Voutetakis², D. Tzovaras¹
¹Information Technologies Institute, Centre for Research and Technology – Hellas (CERTH/ITI), 57001 Thessaloniki, Greece
²Chemical Process and Energy Resources Institute, Centre for Research and Technology – Hellas (CERTH/CPERI), 57001 Thessaloniki, Greece
E-mail: {elithoxo,thanvaf,krinidis,djoannid,dimitrios.tzovaras}@iti.gr, {cziogou,paris}@cperi.certh.gr

Abstract—The determination of abnormal behavior at process industries gains increasing interest as strict regulations and highly competitive operation conditions are regularly applied at the process systems. A synergistic approach in exploring the behavior of industrial processes is proposed, targeting at the discovery of patterns and implement fault detection (malfunction) diagnosis. The patterns are based on highly correlated time series. The concept is based on the fact that if independent time series are combined based on rules, we can extract scenarios of functional and non-functional situations so as to monitor hazardous procedures occurring in workplaces. The selected methods combine and apply actions on historically stored, experimental data from a chemical pilot plant, located at CERTH/CPERI. The implementation of the clustering and classification methods showed promising results of determining with great accuracy (97%) the potential abnormal situations.

Keywords—data mining; elbow method; silhouette method; one class svm method; fault detection; industrial environment;

I. INTRODUCTION

In the past decades a significant effort and resources have been invested by the industrial sector to analyze the growing amount of large data sets that are produced from daily procedures, in order to model machine behavior(s) that operate in continuous, batch or discrete mode. Furthermore, as industries need to operate in a highly dynamic and competitive environment with strict targets to improve production and at the same time to reduce operational cost and save resources, the need for maximization uptime and minimization of unplanned maintenance procedures is imperative. In that context, the detection and diagnosis of malfunctions, along with an early warning notification mechanism is of great interest for every industrial process. Numerous studies have been performed in the field of fault detection in industrial environments by applying either stand-alone data analytic methods or combination of them. The main goal of this work is the ability to monitor real time systems with the concurrent analysis of the imported data, acquiring information that may lead to repetitive patterns (or clusters) that point to either normal function or malfunction of the industrial process.

The term fault detection (or anomaly) refers to a problem of finding patterns in data that do not conform to expected normal behavior. The importance of fault detection is due to the fact that abnormalities in data can, sometimes, point to significant (and often critical) actionable information. Most of the existing fault detection techniques confront deal with the specified problem, based on their application domain and not in a generic form. In literature, one can find many approaches that deal with on fault detection, such as artificial intelligence (i.e. classification [1]), clustering approach [2] and statistical approach [3].

In the context of the industrial processes applications, this work considers a continuous chemical process, with emphasis on the behavior of the involved chemical reactors. Our objective is to apply data mining techniques for the autonomous analysis of large datasets to improve process monitoring, process status inference (soft sensors), detection of abnormal situations (potential faults) and their diagnosis in conjunction with the improvement of the process understanding through the discovery of correlations between process control loops.

A. Visual Analytics

As large amount of data are produced at industrial environments, there is an imperative need for visual representation and information interpretation that can be facilitated by Visual Analytics (VA) methods. In general, VA is an iterative process that involves information gathering and visualization [4, 5] data preprocessing, knowledge representation, interaction, automatic data analysis and finally decision making [6]. VA can be described as “the science of analytical reasoning facilitated by interactive visual interfaces” [7]. In the field of industrial processes, a combination of various methods is implemented for malfunction detection and incident detection [8, 9], such as the use of the Tennessee Eastman process [10]. Alternative methods, such as vibration analysis have been developed to determine if a machine is functioning correctly or not [11] or the use of the Taguchi’s method to explore potential multi-level state of damage in gearboxes [12]. Furthermore, fault detection methods are highly important at areas like civil protection [13], transportation systems [14], internet security [15], and big data analysis [16] and at the field of facilities management [17, 18].

B. Data analytics

Besides the visual analytics tools that can be used for incident detection and abnormal operation representation, data
analytics tools and algorithms have received great attention over the past few years, which is indicated by respective state-of-the-art review studies [19]. The respective methods which are developed for knowledge discovery using data analytics, meaningful clusters and anomaly identification, are applied to a variety of fields such as incident detection for traffic conditions and management, meteorology and earth sciences [20, 21], communication and mobile networks [22, 23], applied economics [24] urban data [25] and ecometrics [26, 27]. The aforementioned indicative cases demonstrate clearly a significant need for further research on integrated knowledge discovery concepts to generate versatile process-centric data mining applications that can be adapted to heterogeneous industrial environments and beyond.

The proposed approach is a systematic design and development of a knowledge discovery framework, which is based on the pattern analysis through clustering methods. Furthermore, the development of a data analytics and visualization application is performed, which is used for the active notification of, involved process operators and supervisors in industrial environments for abnormal operation and increased fault potential. The development of the knowledge discovery framework is based on shop floor knowledge and on actual experimental process data from a pilot plant of a chemical process. The challenge is, firstly to discover hidden insights among the features of the dataset at any level and whether they can be clustered and secondly, to use this output for a training procedure in order to point to successful and useful knowledge discovery. Subsequently, this knowledge may reveal a repetitive behavior every time a malfunction happens, and if so a fault detection procedure will be modelled afterwards.

This paper is organized as follows. In Section II, the data analytic methods used for the pre-processing and clustering of datasets are described. Section III presents the problem statement and the method for determining the number of appropriate clusters for the case into consideration. Section IV analyses the results from the clustering and classification of data and their representation through a VA approach. Finally, in Section V conclusions and future work are discussed.

II. ANALYSIS OF METHODS FOR DATA CLUSTERING AND DATA MINING

As stated above, when large amount of data are present in a specific industrial environment, it is difficult to interpret or cluster them manually or with heuristic methods. Therefore, the use of data mining techniques at the aforementioned fields is facilitated. The importance is to understand the kind of data which are analyzed and discover interesting knowledge and patterns by applying the most appropriate methods. Usually the main scope of data mining is to understand if there are groups among them and if so, to find these groups and explain their interrelation.

A variety of clustering methods exist which indicates that the selection of the suitable one for a particular industrial process is important. The initial step is to determine the optimal number for the data clusters and train a model in order to recognize malfunctions. Overall in this work, two methods are chosen in order to identify the right number of clusters, the Elbow method and the Silhouette method. Subsequently, according to the nature of the problem that is dealt in this work, the One Class SVM method is used to detect outliers in the dataset and discover patterns or models that lead to abnormal operation detection.

A. Elbow Method

The Elbow method determines the number of clusters by checking the percentage of explained variance. During the data analysis, the appropriate number of selected clusters, is based on the variance explained by existing ones and the variance added by a new cluster. If the newly added cluster does not provide much better modelling of the data, the proper number of cluster to be chosen is the one prior to this addition. If the percentage of variance is plotted against the number of clusters, it is noticeable that in the beginning, there is a lot of variance explained, but at some point the marginal gain will drop because of the abatement of the data explained and the plot will gain an angle. This angle is known as the “Elbow criterion”. Further to this, the basic idea behind the method is to define clusters such that the total intra-cluster variation (known as total within-cluster variation or total within-cluster sum of square) is minimized:

$$\min (\sum_{k=1}^{n} W(C_k))$$ (1)

where n is the total number of clusters, $C_k$ is the kth cluster and $W(C_k)$ is the within-cluster variation. The total Within-cluster Sum of Square, denoted hereafter as WSS, measures the compactness of the clustering and we want it to be as small as possible. Due to the ambiguity of the Elbow angle that may occur in some situations, the Silhouette method is also used for both identification and verification of the choice.

B. Silhouette Method

Another useful criterion for deciding about the number of clusters is the average Silhouette method of the data. This is a measure of how closely the Silhouette of a data instance is to the data within its cluster and how loosely it is to the data of the neighboring clusters. The data are correctly formatted to clusters when the average distance of the clusters is the lowest respectively to the data. It is a criterion that measures the quality of the formulated clusters and depends on any distance metric like Euclidean or Manhattan. Consider that the number n of clusters has already been decided via the Elbow method. For each data instance i in a cluster, let $a(i)$ be the average dissimilarity with the rest of the data. Thus, if $a(i)$ is small the data is clustered correctly otherwise they are not. Afterwards, the average dissimilarity of this data instance to a different cluster is defined and this is repeated until it is computed for all clusters. We assume that $b(i)$ is the lowest average dissimilarity of the data instance for any cluster that it is not part of, this cluster is the neighboring as it is the closest to the data instance apart from the one it belongs. The Silhouette is defined as follows:

$$s(i) = \frac{b(i)-a(i)}{\max[a(i),b(i)]}$$ (2)

which concludes into:

...
Thus, $-1 \leq s(i) \leq 1$. The Silhouette method scores between $-1$ to $1$ range. This means that when Silhouette is near to 1 the data are in the correct clusters while when it is near $-1$ data are not clustered correctly. Moreover, it is possible to re-scale data so that the Silhouette method can be maximized at the optimal number of clusters. Consequently, in order to evaluate if data are clustered in the best possible way, the average $s(i)$ of all of them is calculated. Thus, Silhouette plots may be used to determine the natural number of clusters within a dataset. Meanwhile, having defined that data can be clustered and knowing the aim is to implement fault detection, One-Class SVM is applied in order to train the application on normal and case that a malfunction occurred.

C. One-Class SVM method

The One-Class SVM method separates the training data from the origin and finds the maximal margin hyperplane by iteratively mapping input data into a high dimensional feature space via a kernel. This method can be viewed as a typical two class SVM algorithm which puts all the training data in class one and the rest are thought to be outliers or class two. The hyperplane corresponds to the following classification rule:

$$f(x) = (w, x) + b$$

where $w$ is the normal vector and $b$ is a bias term. It is represented by $w^T x + b = 0$, where $w \in \mathbb{F}$ and $b \in \mathbb{R}$. The main scope though, of the method is to solve an optimization problem in order to find the rule that has the maximal geometric margin. So labeling a subset $x$ of a dataset as normal occurs if $f(x) < 0$ and abnormal otherwise, fact that arises from classification rule. Actually, what happens is an exchange of maximizing either the distance of the hyperplane from the origin or the number of training data points that form the area which is separated from the origin by the hyperplane.

In order to avoid overfitting, variable $\xi_i$ is used for letting some data points lie with the margin. The optimization solves a dual quadratic programming problem:

$$\min_{w, \xi, \rho} \frac{1}{2} ||w||^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i - \rho$$

which subject to:

$$\begin{align*}
(w \cdot \varphi(x_i)) &\geq \rho - \xi_i, \text{for all } i = 1, ..., n \\
\xi_i &\geq 0
\end{align*}$$

At Eq. (5), the parameter $v$ characterizes the solution by setting an upper bound on the fraction of outliers and a lower bound on the number of training data which are used as support vectors. Additionally, with the use of Lagrange techniques to set weights to the vectors and a kernel function for the dot product calculations, the function becomes:

$$f(x) = sgn((w \cdot \varphi(x_i)) - \rho = sgn(\sum_{i=1}^{n} a_i K(x, x_i) - \rho)$$

Thus, the method forms a hyperplane which depends on $w$ and $\rho$ and has a maximal distance from the origin in the feature space. The kernel function $K(x, x_i)$ projects the input vectors into the feature space with non-linear decision boundaries and suggests that a feature map is $\varphi : \mathbb{X} \rightarrow \mathbb{R}^N$, where $\varphi$ is the projection of input X to a multi-dimensional feature space, then the kernel is defined as follows:

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle$$

Three different kernels are usually applied in the One-Class SVM, the linear, the polynomial and the Gaussian Radial Base Function (RBF), which are defined as:

- **Linear Kernel**: $K(x, y) = (x \cdot y)$
- **Polynomial Kernel**: $K(x, y) = (x \cdot y + 1)^d$, where $d$ is the degree of the polynomial
- **RBF Kernel**: $K(x, y) = e^{-||x-y||^2/(2 \sigma^2)}$, where $\sigma^2$ is the variance

In general, the selection of the suitable kernel for the One-Class SVM is done by using prior knowledge of invariances or in the absence of experts it can be automated selected by optimizing a cross-validation based model selection.

III. DESIGN AND ANALYSIS OF THE KNOWLEDGE DISCOVERY APPLICATION

This work focus to the design, development and implementation of a data analytics/visual analytics application to explore the large amount of data which is produced by a chemical process pilot plant along with the interpretation of the behavior of the selected areas. The steps for the development of the knowledge discovery framework that targets the malfunction detection and classification are: problem definition, data preparation, data analysis techniques, visual analytics, data evaluation and results interpretation.

A. Analysis of Heating Zones of a Chemical Reactor

Monitoring and diagnosis are important areas in process systems where data mining analysis can be effectively applied. An indicative example is the study the long-term performance deterioration or the abnormal behavior against predefined profile of operation in processes, using historical data in conjunction with online analysis to identify potential abnormal situations. The produced data can be analyzed to monitor the process at different stages. When monitoring processes, the data models created with data mining tools can determine whether or not the current state can generate satisfactory outcomes or in case abnormalities are detected, then provide respective notification to the operators.

In order to evaluate the data mining techniques, a chemical process system that is situated at the premises of CERTH/CPERI is used as a pilot case. During the evolution of pilot plant experiments, a variety of sensors produce large amount of data that are related to process variables such as temperatures, gas flows and steam pressure. In this work, the behavior of the heating zones of a chemical reactor is studied.
In principle, the temperature conditions of a reactor are maintained by a set of heating zones that are affected by the dynamically evolving exothermic or endothermic reactions. For each heating zone there is a measured input variable, the temperature, and an output variable, which is the percentage of operation of the heating resistance. In order to maintain the heating zone to a desired temperature set point, a controller is used that defines the percentage of operation of the heating resistance according to the measured temperature. A PID controller (Proportional–Integral–Derivative) is used as a control loop feedback mechanism and continuously calculates an error value as the difference between the measured temperature and the desired set point. The controller generates actions to minimize the error over time by adjusting the control variable, which is the power supplied to the heating element.

B. Problem Definition and Procedural Steps

The motivation for developing a knowledge discovery framework that relies on data mining methods and supports decision making procedures arose from the fact that the malfunctions during the operation of process systems are translated into downtime cost and loss of valuable material and products. The idea is to study the behavior of large data sets, which represent values from the process variables, and to detect unusual behavior in order to provide appropriate notification and warning for potential malfunction at specific area of the plant. This problem can be approached by different ways [28, 29] based on the nature of the data giving which may explain a set of certain observations.

![Flowchart](chart.png)

Fig. 1. Procedure for knowledge discovery framework for malfunction diagnosis.

In this work, observations are derived by the analysis of the temperatures and the respective controllers that produce a tuple of time series of independent variables. The objective is to discover clusters that may exist in these data sets, which are produced during the operation of the process system. These clusters refer to functional and non-functional state of the heating zone. Fig. 1 shows a general view of the procedure, which has been designed in order to develop the knowledge discovery framework for malfunction diagnosis for these time series.

In order to test the developed methodology, a scenario derived by experimental data with and without malfunctions is selected. During the nominal operation of the process plant a malfunction of a heating resistance occurred that forced the process operators to stop the experiment in order for the technical team to perform the necessary unplanned maintenance activities to repair the resistance. The respective safety system prevented any potential accident, but nonetheless the experiment had to stop and could only start after the repair of the heating zone.

Although the process operator are able to monitor each thermocouples, due to the large amount of the process variables (> 900), it is difficult to manually monitor the correlations between the temperature and the respective controller response. Therefore, the main scope of the proposed framework is to derive automatically information from the large data sets and transform it into knowledge in an understandable structure for further use.

C. Data Preparation-Selection

In this work, data are stored in a warehouse that has been developed in order to process selected data from the Process Information Management System (PIMS) of the plant floor. The aforementioned warehouse relies on a Common Information Model. A subset of the process plant variables is selected, related to 129 sensors of the heating zones, which send the respective measurements of a sampling interval of 3 sec. After studying the large data set of historical data containing days with malfunctions, some correlations arose between the response of the controller and more than one thermocouple. In Fig. 2 the correlation values of all possible combinations pairs of the 210 thermocouples, is given. The pairs that are most correlated or anti-correlated are those that approaches value 1 and value −1, respectively and are the ones that are studied in this work.

![Graph](graph.png)

Fig. 2. Correlation values (r) of all combinations pairs of thermocouple.

D. Data Analysis for the Definition of the Number of Clusters

After the definition of the problem and data preparation, the definition of the number of the clusters follows. The Elbow method is applied to the data in order to retrieve the most
appropriate number of clusters and understand the information which is extracted from this grouping.

This method checks the effect of the variance of the data against the number of clusters that are used to group them. In this work, the method performs a clustering for a range of n-clusters, from 1 to 10, and measures the Sum of Squared Errors, denoted hereafter as SSE. After plotting the points of the SSE, which reflect the structure of an “arm”, the point where the “elbow” appears is usually the appropriate number of clusters that is suggested to be applied to the data.

The aforementioned analysis, using the Elbow method and the Silhouette metric, suggest that the appropriate number of clusters is 3.

IV. RESULTS ANALYSIS FOR THE MALFUNCTION DETECTION AT THE CHEMICAL PROCESS PLANT

In this work, three clusters are used for distinguishing the status of the process plant which refers to identification of the malfunction incident. In term of operational interpretation, the clusters refer to the values during the nominal operation of the pilot plant, the data when the malfunction appears and the data close to abnormal operation before the incident. The idea now is to see if a clustering method is able to separate the data correctly to the three clusters combined with appropriate labels according to their physical meaning. The problem defines these labels according to an empirical study that took place at the plant floor where the expected values and the actual values are defined during a single day scenario based on the acquired experimental data. The expected values are the thresholds that the process engineers have determined based on the operating conditions of the chemical reactor and the safety of the equipment. The formula defines three labels according to the difference between the expected values and the actual ones.

A. Data Clustering – Training and Evaluation

In order to evaluate the detection of the incident, the One-class SVM method is applied, as it seems to be one of the most efficient machine learning techniques for the fault detection. Overall, the One-Class SVM method is designed to estimate the support of a high-dimensional distribution in order to define how well or not the data are fitted through this method. Interestingly, after the extraction of the number of clusters in the data, the idea is to see how near or far to the frontier of normal function (operation), the rest of the data will be plotted. According to the results from the Elbow and Silhouette methods, if the model is trained with the values of the cluster that is considered as normal function, the frontier can determine that when the values are outside of it, but quite near to it, may indicate an impending incident, whereas those far away denote the actual incident. The One-class SVM takes a small percentage of the training data as support vectors and tries to fit, if possible, a significant amount of the rest of these data, in a “frontier of trust” using a polynomial kernel. Because of the linear dependence of the data, the specific kernel is selected for the training. All data are normalized between -1 and 1.

In our case, the method trained the model successfully by ranking 0% in outliers. All data used for normal function are correctly clustered as 1 using polynomial kernel of degree 3. The frontier that separates space in two classes shows the limit where an observation is considered normal or fault. By plotting data characterized as fault in this research, they are correctly classified as −1 outside the “trusted area”. Some areas though are formed which according to shop floor knowledge, verify the idea of three clusters in this issue. In the fault cluster the classified observations consist of two smaller clusters. So, based on the operation conditions and the behavior of the pilot plant, the data modeling can be implemented based on the following rules for labeling:

Assume that \( d = Actual\ Value - Threshold \)
If $-20 \leq d \leq 20$ the observation is labelled as 1 (Normal Functionality)

If $d \leq -20$ or $d \geq 20$ the observation is labelled as -1 (Malfunction)

The One-Class SVM needs only the training data to perform clustering, so the rest of the labelled data are kept for the validation of the entire approach. Fig. 5 shows a day with an incident at a defective heating zone (using pairs of thermocouple, controller output values), where the training dataset consists of observations from the normal function as defined from the One-Class SVM method. Also, the results of the training procedure are presented and the formation of the three types of clusters. The observations that form the cluster before the incident are classified correctly up to 99% while the actual incident observations are classified in a level of 93% correctly.

Fig. 5. Normal functionality and malfunction detection: training of One-Class SVM method.

After the training of the model, the validation of the methods is performed using a new set of observations of normal function, in order to see whether the method is going to cluster them correctly. The training procedure has created the frontier so in Fig. 6 the received data from a normal functioning time interval, are classified correctly in the area that is characterized as trusted from the model. The success percentage of the classification of the new normal observations is 86% while the rest of them are classified closely to the frontier but as fault.

In Fig. 6b, abnormal function is identified outside the frontier and by enlisting correctly the malfunctioning values, the method succeeds in detecting the anomaly in data. Results showed 97% success in classifying correctly fault observations, so anomaly detection for this kind of data is accomplished.

B. Visual Analytics

Besides the data analytics, this work deals with the VA part of the knowledge discovery framework in order to provide a monitoring tool at the plant floor. Therefore, the next step involves the extraction of knowledge from historical data and real time data and their appropriate representation. A graphical representation with nodes is used to provide an overview of the operation status of the chemical process into consideration (see Fig. 7).

Fig. 7. BIOCAT has five subsystems and each one has sensors that change color according to received values.

The chemical process, named BIOCAT, consists of five different subsystems, the Regenerator, the Riser, the Stripper-Liquid Products, the Injector and the Lifeline. The subsystems have various thermocouples and the knowledge discovery framework sends values to the respective nodes that represent the overall status of the involved equipment.

Each node represents a unique sensor which changes color depending on the status value that is received from the evaluation of the clustering and malfunction analytics method. When the node starts getting color close to red, then the user needs to be notified accordingly in order to pay attention to the specific area as a potential incident might occur. A specific time interval is required to pass with the node to remain at the red representation, in order to consider that a malfunction has occurred. Moreover, this aims to give an abstract overview, allowing the analyst, designer or engineer to identify problematic aspects in terms of performance and further focus on in them towards better understanding the source of any
performance degradation problem through comparative analysis and correlation. In order to have the desired results, the aforementioned techniques are applied for the development of the specific VA interface.

C. Evaluation and Interpretation

The main scope of this work is to understand if patterns and useful information exist at the large data sets for troubleshooting the case of anomaly detection. It must be noted that the development of the knowledge discovery framework is an iterative procedure that targets to the extraction of valid and initially unknown information from a large amount of data and provide empirical data models with feedback. Usually the root cause of faulty functioning equipment or systems is not an easy task. Therefore, the results were discussed with domain engineers and data experts to find a possible model and solve the problem. According to the analysis results and discussion with the domain experts, values of historical incidents reveal the existence of a cluster that is neither faulty nor normal function. This clustering approach may provide useful early notification or warning mechanism in order to check the thermocouples that are classified in it for an adequate time period. The evaluation of the results and the application of the data mining and visual analytics methods showed that the pilot plant data can be processed accordingly and help process operators and supervisors to be notified about fault diagnostics and as an extension the prevention of incidents.

V. CONCLUSIONS – FUTURE WORK

The proposed approach that was presented in the current work, targeted at the extraction of information from large data sets in order to facilitate the diagnosis and prevention of malfunctions at an industrial environment. The developed schema has been applied on historical experimental data and the annotation of functional and non-functional states of the chemical process into consideration has been complemented by the experience of domain experts.

The evaluation of the deployed framework using experimental data from the specific chemical process, demonstrated that the approach was able to correctly distinguish data during normal operation and when a malfunction occurred, up to 86% and 97%, respectively. Furthermore, the trained model was able to identify the cluster which is related to the situation of an imminent incident up to 99% and it was observed that the distance from the learned frontier was smaller than the respective data related to the malfunction case. The experimental results were analyzed by process engineers and the proposed tools have a promising potential for further use as they could provide a notification or early warning mechanism for imminent incidents.

The evaluation of the aforementioned cases provide a clear indication that knowledge discovery using data mining and clustering are useful tools for the identification of the problem of malfunction diagnosis in industrial processes. Therefore, as future work, the utilization of alternative machine learning techniques, such as Decision Trees, Random Forest, AdaBoost, etc., is a promising potential in exploring the behavior of the heating zones of chemical reactors and other subsystems or control loops of process systems.

Based on that, we plan to explore further the problem of malfunction diagnosis on industrial environments utilizing not only different machine learning schemas but also trend analysis so as to develop a robust and concrete inference engine. Moreover, we aim to apply the proposed off-line data analysis, along with the future modifications, in the real time environment of the pilot plant.

ACKNOWLEDGMENT

This work has been partially supported by the European Commission through the project HORIZON 2020- INNOVATION ACTIONS (IA)-636302-SATISFACTORY.

REFERENCES


