DE GRUYTER OPEN Acta Univ. Sapientiae, Informatica 9, 2 (2017) 101–118

DOI: 10.1515/ausi-2017-0007

Anomaly detection techniques in cyber-physical systems

Gheorghe SEBESTYEN
Technical University of Cluj-Napoca, Romania
email: Gheorghe.Sebestyen@cs.utcluj.ro

Anca HANGAN
Technical University of Cluj-Napoca, Romania
email: Anca.Hangan@cs.utcluj.ro

Abstract. Nowadays, when multiple aspects of our life depend on complex cyber-physical systems, automated anomaly detection, prevention and handling is a critical issue that influence our security and quality of life. Recent catastrophic events showed that manual (human-based) handling of anomalies in complex systems is not recommended, automatic and intelligent handling being the proper approach. This paper presents, through a number of case studies, the challenges and possible solutions for implementing computer-based anomaly detection systems.

1 Introduction

Anomaly detection in physical processes (form very simple ones like an electric motor toward very complex industrial infrastructures) is not a new task; it is part of the operating and maintenance procedure of that system. Because of the multitude of anomaly sources and consequent system behaviors this task was traditionally left to the experience and intuition of a human operator. But in today's complex cyber-physical systems with thousands of process variables

Computing Classification System 1998: B.3.4, B.8.1
Mathematics Subject Classification 2010: 68M15
Key words and phrases: anomaly detection, outliers detection, cyber-physical systems, time series, spatio-temporal data

involved and multiple automatic control loops the ability of a human operator to identify an abnormal behavior or state is overwhelming. Sometimes the required reaction time to a given event is much under the typical reaction time of a human (which is usually greater than 0.1 s).

There are a number of examples of catastrophic events caused by the fact that an abnormal system behavior was not properly identified and handled. For instance, recently (Aug. 2016) an explosion could have been avoided at the Petromidia petroleum processing plant (in Romania) if the human operators wouldn't have ignored a sequence of alerts that signaled a gas leakage [13]. But the real problem was that more than 1.6 million alerts were generated by the automated system in the last 3 days previous to the explosion, probably a lot of them being false alerts. In front of such a huge number of alerts a human cannot identify and classify the anomalies and threats at their correct risk level.

Therefore automated algorithms and methods are needed to identify and handle in real-time critical system anomalies. But implementing efficient anomaly detection methods is not a trivial task. For example, for a person is rather simple to say that something is wrong with his/her car based just on the sound generated by that car and a specialist can even tell the component that cause the trouble. Transposing such an intuitive detection into an algorithm or automated method is not a straightforward task.

The difficulty starts with the definition of an abnormal behavior or simply of an anomaly. It continues with the multitude of possible anomalies and sources of anomalies. An anomaly may be caused by accidental (non-malicious) causes such as: a communication error, faulty equipment or measuring device, a noisy signal and significant environmental changes; it may also be caused by intentional (malicious) actions, such as: a virus, an intruder or a theorist. There are examples of cybernetic attacks specially designed for very critical cyber-physical and embedded systems (e.g. Stuxnet, Duqu).

It is generally accepted that an anomaly is a deviation from a normal state or behavior; therefore it is important to identify a normal state (or states) as a discriminant for identifying anomalies. As it will be showed in the case studies, most of the proposed anomaly detection mechanisms are trying to identify a number of relevant features of the analyzed system that allows making the difference between normal and abnormal behavior.

Due to abnormal system behavior, monitoring data sets include outliers. The term "outlier" was originally used in the field of statistics and it is [1] defined as an observation that is inconsistent with the set of data it belongs to. Even if they are not quite equivalent, in many cases the terms "outlier detection" and "anomaly detection" are used with the same meaning. In our view outlier detection is more an off-line process, while anomaly detection is an on-line one.

As presented in a very good survey on anomaly detection methods [2] there are different forms of anomalies, different anomaly sources, different types of systems (system behaviors) and different application domains. Therefore there is a very wide range of methods used for this purpose, "borrowed" from multiple domains, such as: statistics, data mining, machine learning, information theory, signal processing and spectral analysis, etc.

The goal of this paper is to analyze through some examples those methods that are best fitted for the cyber-physical domain. Typical for this domain is the use of sensorial networks and sensorial data, the need for on-line (real-time) analysis and detection and the presence of multiple correlations between the acquired data. As shown in the next chapters, the common feature for the methods applied in different case studies is the identification of an anomaly as a value or a state that breaks the previously detected or learned correlation rules.

This paper is a retrospective survey of our manifold research in the area of anomaly detection applied in different domains and for various purposes.

The rest of the paper is structured as follows: the next section presents some basic concepts and related research in the field of anomaly detection, specific for cyber-physical systems. Section 3 tries to classify the different conceptual approaches for anomaly detection and analyze the possibility to adapt a given method to the specificity of physical systems. The next sections present a number of case studies for different types of anomaly sources and system types. These sections reflect some of our previous results in different areas. The last section presents our conclusions and some future research possibilities.

2 Related work

There are several recent survey papers that try to organize and classify the large amount of research work that has been conducted in the field of outlier or anomaly detection, while highlighting the research issues that still need attention [12, 6, 4, 11].

The authors in [12] identify three large types of outlier detection problems based on outlier sources: fault detection in case of noise and defects, event detection in case of multi-variable systems and intrusion detection in case of malicious attacks. One of the main challenges of outlier detection in sensor networks is in fact identifying the source of outlier data, since traditional techniques fail to distinguish between errors and events. Other important challenges identified in [12] are related to the scalability and the computational complexity of the detection techniques. The authors classify the outlier detection methods in statistical-based approaches, nearest-neighbor based approaches, clustering-based, classification-based approaches and spectral decomposition-based approaches. They point out that the first two classes can't handle multivariate data sets and the other, more complex methods can't be easily used for large scale sensor networks because of their large resource requirements or computational complexity.

The authors of [6] identify the requirements for an efficient and effective anomaly detection model that include five items: reduction of data, online detection, distributed detection, adaptive detection and correlation exploitation. They point out that current anomaly detection models have important limitations such as the failure of adaptability in dynamic environments, not taking into account spatial and temporal correlations between data and the absence of automated parameter tuning.

Other surveys [4, 11] extensively cover outlier detection techniques that are used for the detection of malicious attacks. The authors in [4] mainly cover the problem of data injections in sensor networks. As they classify the techniques used for the detection of anomalies, they emphasize the importance of attribute, temporal and spatial correlation in solving the problem of multiple compromised sensors that produce anomalous values in a coordinated fashion. In [11], the authors make a classification of security threats in sensor networks and of the outlier detection methods used. They conclude that data mining and computational intelligence based schemes are the strongest in terms of detection generality as long as the adequate attributes are selected. Finally, they identify some potential research areas such as modeling the problem of anomaly detection, attribute selection and the development of a uniform performance evaluation standard.

In this research context, our paper tries to give a more pragmatic approach to the anomaly detection problem. Through the case studies we show that the key for any anomaly detection method is to find the set of features that discriminate between normal and abnormal system states, process variable values or events. It is also important to find correlations between process variables that are broken in case of an anomaly.

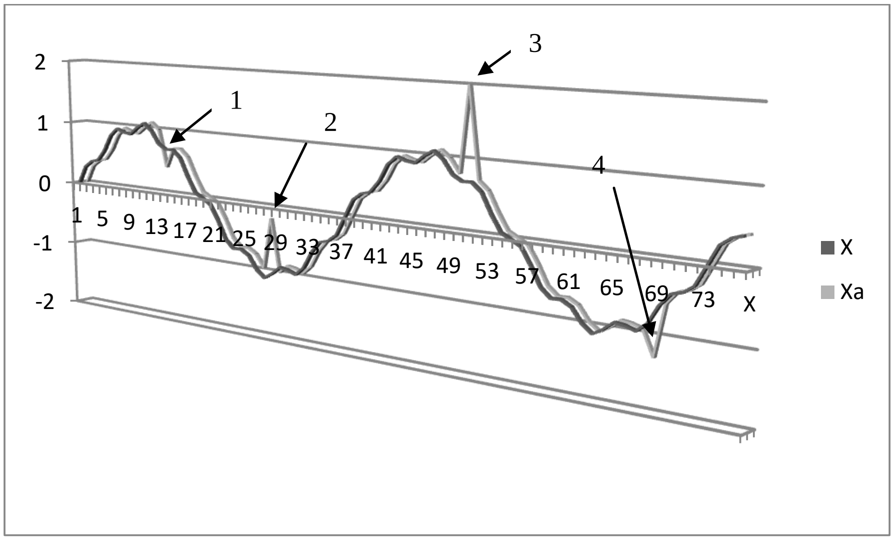

Figure 1: Anomalies in time series (X : original signal, Xa : signal with singular anomalies).

3 Anomaly detection techniques

An important group of anomalies are singular values that do not fit with the rest of the acquired data. In time series these outliers can be seen as values that do not follow the continuous shape of a variable graph. Some values, which are outside of a normal variation range (e.g. dots 3 and 4 in Figure 1), can be detected if a minimum and maximum value is set or detected on a training set. Other values are in the normal range but it is still obvious for a human eye that something is wrong (e.g. dots 1 and 2 in Figure 1). These outlier values can be detected with linear and parabolic prediction or through autocorrelation techniques (see more details in next chapter, case study “a”).

Also singular anomalies may be detected in spatially distributed variables. For instance, in environmental monitoring systems, values (e.g. temperature, pressure, humidity) measured in a small vicinity tend to be similar or at least correlated somehow. In such systems (see Figure 2) an outlier is a value that does not fit with the spatial curves of the neighbor values. Linear approximation and spatial correlation techniques may be used for detection. Sometimes time and spatial correlation may be combined for more accurate anomaly detection.

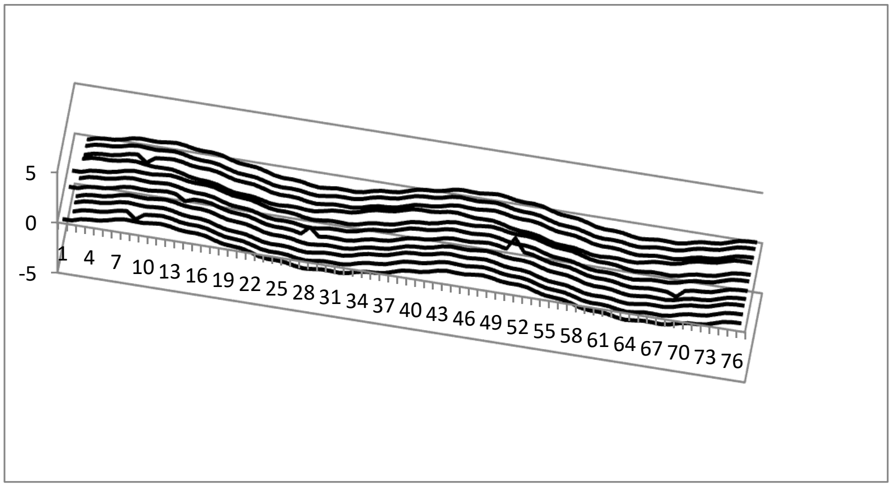

Figure 2: Anomalies in spatially distributed values.

In a cyber-physical system there are multiple functional correlations between process variables, which can be used for anomaly detection. These functional dependencies may be theoretically deducted from the physical and chemical laws that govern that process or they may be determined experimentally from the measured data sets. Figure 3 shows 5 process variables and Table 1 presents computed correlations between some pairs of variables. Through correlation values we can establish that variable v is mostly correlated with x and z variables and less correlated with y and u. In this case there is a functional relation between v, x and z, which may be exploited for anomaly detection.

Correlation	x and Y	x and v	z and v	y and V	u and v
Values	-0.66	0.91	0.89	-0.64	0.19

Table 1: Correlations between pairs of variables.

Another category of anomalies (beside singular ones) are those that change the typical shape of a signal. In this case the allowed variation domain or the “continuity” feature of the graph are not violated and therefore other techniques must be applied, techniques that recognize the normal and abnormal shape of the signal. Here, pattern recognition and classification methods are used. For instance, a doctor can recognize a given heart disease based on the specific normal and abnormal ECG signals. A pattern recognition tool (e.g.

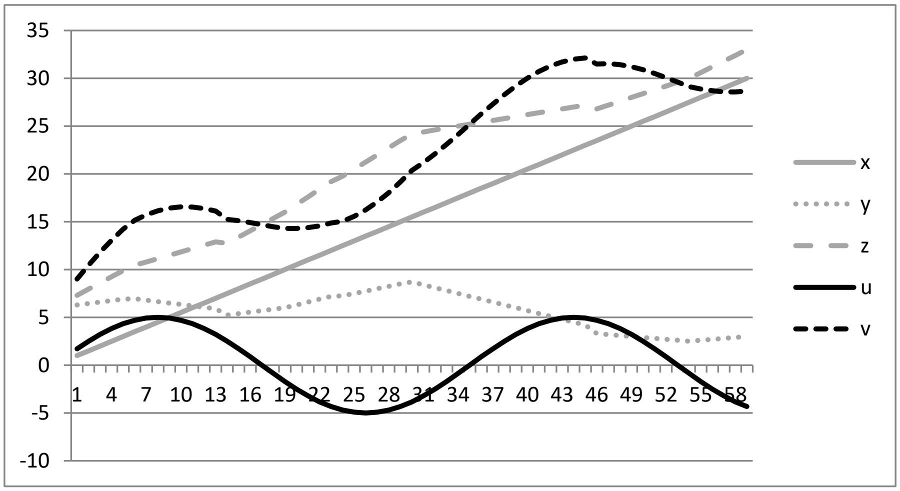

Figure 3: Functional correlations between variables.

neural network, decision tree) is trained with normal and abnormal ECG signal shapes. But in cyber-physical a system, generating abnormal signal shapes is not a trivial task (it may even destroy the system) and there are many abnormal behaviors, most of them not predictable from the design phase.

An interesting approach in this area is to classify in simple terms (e.g. letters or codes) the different slopes of a signal and then identify a normal or abnormal behavior based on the sequence and duration of codes. We used this approach for identifying road anomalies (see case study "d") and also abnormal behavior of elderly persons [8]. Because the human behavior is rather complex, with multiple possible choices, normality was hard to define. Hidden Markov chains were trained in order to classify normal and abnormal behavior.

4 Case studies of anomaly detection

This chapter gathers a number of relevant cases regarding anomaly detection methods developed for different purposes. In every case we analyze the main goal of the detection, possible methods and expected outcomes.

4.1 Anomaly detection in sensorial networks

In case of sensorial networks most of the methods used [2] try to exploit the existing correlations between the data acquired from sensors. Usually there are three types of correlations that may be identified in such systems:

- Time correlation or correlation of a sensor with itself;
- Spatial correlation or correlation between a sensor and its neighbors;
- Functional correlation or a correlation imposed by the functional relations between components of a complex system.

The first kind of correlation is specific for process variables that have a quasi-continuous evolution in time and their future behavior can be predicted from their past values. In this case linear prediction and auto-regression techniques can be used. Linear prediction is an easy and fast method that can be implemented even at the intelligent sensor's level. A predicted value $\bar{\mathrm{X}}$ is computed using the last 2 (linear approximation) or 3 samples (parabolic approximation) of the signal. If the difference (ϵ) between the predicted and the last measured value exceeds a given threshold the value is considered a candidate outlier. The threshold can be learned in a training phase as the maximum difference occurred in the training set; the condition is to have a training set without outlier values. Usually the outlier value will be replaced with the predicted one.

The success of this method depends on the granularity of the time sampling (the sampling rate). In order to apply successfully a linear or parabolic approximation the original curve of the signal should be well approximated with line segments or parabolic segments. Our experiments showed that if the sampling frequency is one magnitude (10 times) higher than the highest frequency in the input signal than the approximation error is reasonably small and the error threshold can be kept small. Otherwise, computed differences in the training set will be high and consequently the threshold is too high for a good outlier discriminant. The maximum frequency in the input set can be obtained by applying an FFT on the training set. To avoid false high frequencies generated mainly by noise, the amplitude of the highest frequency in FFT taken into consideration should be a fraction (e.g. 1/10) of the biggest harmonic amplitude. Threshold computation can be done in the initialization phase when no time limits are imposed.

A more computer-intensive method for anomaly detection is through auto-regression. The predicted value is computed as a weighted some of previous samples, as follows:

$$\bar{X}_i[k] = \sum_{j=1}^{N} u_{i,j} X_i[k-j],$$

where $u_{i,j}$ is the weighting coefficient of order j of node i in an auto regression model. Computing u_i coefficients is a computer intensive process, which can be performed only in a device with sufficient computing resources (not on a microcontroller or an intelligent sensor). The coefficients can be computed on the training set, but also on the incoming samples. The window of samples on which the auto-regression model is computed must include a relevant period of time in the evolution of the signal, meaning that the window must include seasonal variations of the time series (variations cause by day-night cycles or season changes). Some programing languages (e.g. the "R" language used by us) have very good library functions for auto-regression and linear modeling coefficients computation.

An outlier value is detected if the difference between the predicted and measured value is higher than a threshold; this threshold can be determined based on the "residuals" of the auto-regression model.

For systems that change their behavior in time the auto regression model should be periodically recomputed on the newly collected data.

Another correlation which may be exploited in sensorial data is the spatial correlation. For instance if there is a set of sensors that are collecting temperature values in a given region it is reasonable to suppose that the values generated by a sensor are in a correlation with the values generated by its neighbors. In this case again a linear model or a regressive model can be computed for each node of the network. Now the predicted value of a node is computed using its neighbors values at the same sampling time or at a lagged time. The lag (or time delay) can be determined experimentally or based on a physical propagation formula (e.g. propagation of temperature gradient in a given environment):

$$\bar{X}_i[k] = \sum_{j=1}^{N} u_{i,j} X_{i,j}[k]$$

where

- $u_{i,j}$ is the weighting coefficient of neighbor j,
- N may be 3 to 8 (for pragmatic reasons),
- $X_{i,j}[k]$ the j-th neighbor of node i.

The vicinity of a node in a sensorial network can be obtained using the geographical position of the nodes in the area (e.g. GPS coordinates). If such information does not exist than proximity of a neighbor can be determined through the radio connectivity between nodes and the amplitude (power) of the radio signal. Of course, a triangulation method would improve the precision of selecting the best neighbor candidates. The number of neighbors may vary from 3 to 8 depending on the available time for computation.

The linear approximation technique can be implemented directly on the sensor nodes. Every node hears (through radio transmission) their neighbor reports and can decide if its value is an outlier. In a similar way the detection can be made by the nodes that aggregate data through the acquisition tree.

The regression model on spatially distributed nodes requires more computing power and can be implemented at the "Access point" node or in a central computer (e.g. server). In the formula that predicts the value of a node at moment "k" we can include the weighted sum of the neighbors' values at the same "k" moment as well as values at one, two or more earlier sampling periods.

$$\bar{X}_i[k] = \sum_{j=1}^{N} u_{i,j,0} X_{i,j}[k] + \sum_{j=1}^{N} u_{i,j,1} X_{i,j}[k-1] + \sum_{j=1}^{N} u_{i,j,2} X_{i,j}[k-2] + \dots$$

where $u_{i,j,l}$ is the weighting coefficient for neighbor j and time delay l.

Functional correlation can be exploited as an alternative for spatial correlation, when the similarity between two variables is in accordance with some functional dependencies between the system's parameters and spatial proximity between two nodes is not relevant. This is the case for a sensor network that collects multiple types of process variable values and there is a correlation between variables in accordance with the physical laws that govern that process. For instance in an electrical energy distribution system the voltage, current, power and energy measurements must be in accordance with the electricity laws (e.g. Kirchhoff's laws).

Functional dependencies between any two process variables can be established on theoretical bases or through an experimental process. In the first case the designer must know a-priory the physical law that govern the process and interconnect the process variables. The system theory shows that finding a true and precise model of a system is not a trivial task and in many cases the multiple external influences (e.g. environmental variations) diverts the system's behavior from the pure theoretical mode. Therefore an experimental approach is more feasible. We can build an experimental model of the system (a process called identification in system theory), or we can compute correlation functions between pears of process variables. For a variable we can consider as its closest neighbors the "N" variables for which the correlation functions are the highest. This computation can be done off-line in the learning phase, based on some previously collected data.

4.2 Pollution detection in rivers using rule-based systems

The detection of abnormal events in environmental monitoring is based on analyzing the values obtained from sensors and the correlations between these values. In the special case of pollution detection in rivers, time, spatial, as well as functional correlations between different parameters have to be taken into consideration.

Several parameters such as temperature, pH, specific conductivity, dissolved oxygen, turbidity and discharge can be used to assess the quality of water. Some of these parameters are measured using sensors (e.g. temperature, pH) and others are computed based on measured values (e.g. discharge is computed based on pressure and river profile measurements [3]). To be able to take advantage of time and spatial correlations, the measurements, acquired from sensors, have to be made continuously in subsequent locations on the river shore for each of these parameters.

Our approach for detecting events while monitoring water quality parameters is a two-step rule based system. In the first step, the parameter values are labeled based on a set of rules that take into consideration time and space correlations between the values measured for each parameter. In the second step, a second rule-based component assesses the functional correlations between several parameters to detect events such as river shore erosion, floods or chemical pollution.

The first step of the rule-based event detection system is focused on the detection of anomalies in the time series of each measured parameter, at each location. These anomalies can be erroneous measurements provided by faulty sensors or values that are outside the accepted value interval, which may signal an event. Labeling rules are different for each parameter, not only because accepted value intervals and correlation rules differ, but also because some parameters' accepted value intervals are variable based on the context in which they are measured (e.g. normal values for water temperature vary based on season). During labeling, it is important to differentiate between erroneous measurements and actual events. This is done by correlating the values measured at subsequent locations. If an event appears at one location, then the measurements downstream for the same parameter will be correlated. More-over, values showing events are time correlated. In case of errors, there are no spatial correlations. A faulty sensor can give unpredictable readings. The labels assigned to the measured or computed values place them in one of the following categories: error, normal, low, high.

Labeled values are passed to the second step rule-based component that will be able to detect actual events based on correlations between several parameters values. For example, river shore erosion may be detected based on high turbidity and high discharge. River shore erosion may signal the risk of floods. If the river shore is near an agricultural land, in the presence of a flood, there is a high risk of nitrate and nitrite pollution. High turbidity is usually detected during and after a rainfall and it causes an increase of temperature and a decrease of dissolved oxygen. This will cause damage to the flora and fauna of the river. Conductivity and pH levels are specific to each water stream due to the soil and geology. Therefore, the change in pH and increased conductivity levels signal the presence of polluting chemicals such as nitrate, phosphate or sodium.

By applying similar water quality assessment rules the second step component will be able to identify various types of events. The performance of event detection is heavily influenced by the quality of preliminary value labeling. An increased spectrum of categories (label types) should improve the assessment process. A partial implementation of this system and its integration with a water monitoring system for Somes River is presented in [10].

4.3 Malicious attack detection using system models

Malicious attacks on cyber-physical systems are another source for anomalies and abnormal behavior of some automatically controlled systems. Before a catastrophic failure happens a number of anomalies may indicate an imminent attack on the system. The goal in this case is to identify the initial signs of an attack and counteract in order to avoid total failure.

One possibility is to use traditional virus and intruder detection methods specific for computer systems. But as showed in [5] cyber-physical systems require specific detection methods that take into account the type of equipment involved (sensors, actuators, regulators, PLCs), the gravity of a malicious attack and the inter-correlation between process variables.

The idea promoted in our research is to try to model the physical process and then simulate different attacks in different points of the infrastructure in order to identify and learn malfunctioning patterns. Then these patterns can be used as discriminants for identifying real attacks.

Two kinds of cyber-physical systems were modeled: a chemical process and an electrical distribution network [5]. In both cases the models allowed us to inject false data at different points and measure their effect upon system variables. It was demonstrated that an efficient attack is not one that try to influence major elements in the infrastructure (they can be detected rather simple in an effective time) but an attack that keep its effect stealthy as much time as possible. In the second case the initial variations are too small for a simple anomaly detector and later when the effects are detectable a significant part of the infrastructure is already under the control of the attacker.

A system model allows an anomaly detector to compute the next predicted value based on the previously measured ones. A maliciously injected value will differ significantly from the predicted one, being a candidate for an anomaly.

Another bases for anomaly detection is an inherent redundancy between measured process variables. The system model gives the inter-conditioning relations between different process variables. For instance in the electrical distribution network example, the sum of the currents going in and out of an intersection must be theoretically zero. In practice, because of the energy loses on the electrical lines an error threshold had to be considered. Similar relations can be found between variables of different types (e.g. power, current and voltage). This kind of anomaly detection can be used not just for malicious attacks but also for malfunctioning components. In the second case (faulty component) the effect tends to be permanent.

The designer of the anomaly detector module can define a set of rules or inter-conditioning relations extracted from the system model. If a precise model of the system does not exist the rules may be formulated based on the experience and intuition of the human operator; in this case Fuzzy relations are preferred.

More difficult is to consider dynamic relations between variables, which are described by differential equations. Dynamic behavior is typical for transitions between more or less stable states of the monitored system. Here an experimentally determined transfer function allows us to write a time dependency between an input and an output variable and then this relation is used for anomaly detection. In system theory this process is called system identification and a number of experimental methods are given for determining the transfer function. Again an error threshold must be considered between the predicted (computed) and measured output variable. The level of the error is influenced by the effect of the noise over the analyzed component (which can be determined in the training phase).

The dynamic behavior should be taken into consideration when the transition periods of the system are more dominant over the stable state periods. In our case, static relations were typical for the electrical distribution network model and dynamic relations for the chemical process.

4.4 Anomaly detection through pattern recognition

A forth direction of our research was to identify an abnormal behavior through pattern recognition techniques. The idea is to collect and learn a number of normal and abnormal behaviors of the system variables (e.g. time variation patterns) and then use them as discriminant for abnormal behavior. There are many methods described in the literature [2] that can be used for pattern recognition (e.g. neural networks, frequency analysis, classification and clustering, SVM, etc.), most of them being time consuming. Our goal was to develop simple methods that can be used for on-line (real-time) anomaly detection and they should be deployed on devices with limited resources.

In this case study [9] our idea was to identify anomalies in the road based on the acceleration signals (on 3 directions) collected from a smart phone placed in a car. The goal was twofold: to identify and locate the holes and speed bumps in a road section and also to give a quality measure of a road section. Through crowd sourcing a realistic and up-to-date map of a given region (e.g. city, highway, etc.) can be obtained and users can be notified about anomalies on the roads they are traveling.

For the first part we implemented a sequence of low and high pass filters that allowed us to discriminate between usual trepidations of the car (caused by low quality roads, acceleration/decelerations, engine rotation, etc.) and variations caused by holes or bumps. Then, with an adaptive threshold we determined a region in the curve as candidate for an anomaly. A neural network was trained to identify different categories of road holes and bumps. As input the neural network considers the order and the sign of the curve slopes and the magnitude and the duration of the abnormal period.

For the second goal (road quality evaluation) a number of features were extracted from the acceleration signals, such as dominant frequencies, component amplitudes and frequency of anomalies. Through calibration we reduced the effect of car speed over the measured signals. More details on this research can be found here [9].

This experiment showed us that simple intuitive rules deployed as a sequence of signal processing procedures (e.g. filters, FFT) allowed us to develop an efficient and real-time road anomaly detection system. Similar techniques can be used for identifying abnormal shapes in the graph of a process variable. Based on our experience we can say as a rule of thumb that if an abnormal shape is recognizable for the human eye than probably a set of signal processing procedures and rules can be implemented in a program that will recognize that shape. This rule applies for anomalies which are a-priori known (as the holes in our experiment).

Anomaly detection based on the signals' shape recognition can be used in cyber-physical systems as well as in many other domains such as: medicine (e.g. ECG complexes, EEG waves, electromyography), electrical and mechanical components maintenance (e.g. early signs of failure), financial processes, meteorology or earth sciences. The methods used are similar but the interpretations are very different.

4.5 Anomaly detection in network traffic

In computer networks the traffic shape and content is very divers because communication applications are run randomly on different computers. Opposed to this case, in networks used for cyber-physical systems the traffic is dominated by periodical data flows. Usually here the data acquisition, processing, storage and visualization are made in a periodical manner and consequently the traffic associated to these activities adopts the same periodicity. Also the order of the activities (tasks) is somehow stable. This quasi-stable state may change if a malicious code tries to infiltrate in the system or if some kind of physical failure occurred and the system reacts with some counter measures. From our point of view both cases can be classified as anomalies.

Based on these observations we can define as an anomaly discriminator a significant change in the pattern of the packages transmitted on the network. The pattern can be identified through the following features:

- The frequencies of different types of packages
- The order of different types of packages
- The lengths of different package types
- Delays between different packages

Through a network sniffer component, in the training phase, the program can identify the types of packages transferred through the network, their periodicity (or their sporadic nature), the typical length of the packages (depending on their type) and the order of the packages. Sometimes these details are a-priori known by the physical-system's designer or by the control systems developer.

Also in some industrial networks (e.g. Profibus, FF, WorldFIP) the traffic pattern is set in the configuration phase and it is strictly imposed through a MAC protocol. Any change in this pattern may be considered an anomaly.

In less restricted networks (e.g. cell industrial networks, Etherbus, CAN, control over Internet) some pattern features can still be detected and transformed into anomaly detection rules. In a training phase a sniffer program can determine all the package types transferred through the network, their length interval (min, max) and repetition frequency. Any significant deviation from normal values is considered candidate for anomaly. Sporadic packages don't have a regular repetition period, but even in this case a minimum frequency can be derived from the physical phenomena or component that initiated it. For instance in a car (on its CAN network), the frequency of packages sent by the rotation sensor placed on the engine cannot exceed the maximum rotation frequency of that engine. Similarly packages reflecting the driver's activity (wheel movement) cannot exceed the reaction time of a human.

In a complex cyber-physical system for reliability and robustness reasons a single anomaly detector is not enough [7]. Multiple detection points must be established in a consistent manner, in different points of the network infrastructure. In [7] we proposed a method for optimal placement of anomaly detectors. The method minimizes a combined cost function that takes into consideration the total coverage of each network node, the transmission overhead and the delays. Further research is needed to identify and express normal and abnormal traffic patterns used by the detector nodes.

5 Conclusions

Analyzing the different cases presented in the paper we can generate a number of rules that may help a developer to select and implement the best anomaly detection solution for a given cyber-physical system. Here are our conclusions:

- Today's cyber-physical systems are becoming so complex and incorporate so many components that a manual (human) anomaly detection is not recommended and in some cases is even impossible;
- Most anomaly detection methods are trying to exploit some regularities or correlations existing between process variables during normal execution;
- The discriminants for detecting anomalies must be built upon a set of signal or system features that mostly change in an abnormal behavior;

- As shown in the case studies, these discriminants are very different, depending on the domain, the source of the anomaly and the complexity of the system;
- In most cases the anomaly detection method must be tolerant with some variations caused by known (e.g. noise) or unknown sources (e.g. Gaussian spread of values);
- In a cyber-physical system multiple anomaly detection points should be spread in the infrastructure and a combination of multiple techniques can coupe better with the multitude of anomaly sources and types.

As future work, based on our previous experiments, we try to develop a platform that will contain a number of anomaly detection tools. This platform will be used by a developer to test the best combination of anomaly detection methods for a given analyzed system. The platform will include facilities for acquiring data from different sources, tools for automatic generation of anomalies and interactive interfaces for flexible result evaluation.

Acknowledgements

The results presented in this paper were obtained with the support of the Technical University of Cluj-Napoca through the research Contract no. 1995/ 12.07.2017, Internal Competition CICDI-2017.

The authors of this paper would like to thank some of our students, whose work was essential for testing some of our ideas and for the implementation of case studies.

References

[1] V. Barnett, T. Lewis, *Outliers in Statistical Data*, New York: John Wiley Sons, 1994. ⇒102

[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, *ACM Computing Surveys* **41,** 3 (2009). ⇒103, 108, 114

[3] P. Deac, M. Muste, O. Creţ, L. Văcariu, H. Hedesiu, A prototype for the continuous and cost-effective measurement of river discharge, *Proc. 2013 19th International Conference on Control Systems and Computer Science (CSCS '13)*, Bucureşti, Romania, 2013, pp. 628–633. ⇒111

[4] V. P. Illiano, E. C. Lupu, Detecting malicious aata injections in wireless sensor networks: a survey, *ACM Computig Surveys* **48,** 2 (2015). ⇒103, 104

[5] I. Kiss, *Security improvement techniques for networked critical infrastructures*, PhD Thesis, Technical University of Cluj-Napoca, Romania, 2016. ⇒112, 113

[6] M. A. Rassam, A. Zainal, M. A. Maarof, Advancements of data anomaly detection research in wireless sensor networks: a survey and open issues, *Sensors* **13,** 8 (2013) 10087–10122. ⇒103, 104

[7] H. Sandor, Gh. Sebestyen, Optimal security design in the Internet of Things, *Digital Forensic and Security (ISDFS)*, 2017. ⇒116

[8] Gh. Sebestyen, I. Stoica, A. Hangan, Human activity recognition and monitoring for elderly people, *Proc. 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj, Romania, 2016, pp. 341–347. ⇒107

[9] Gh. Sebestyen, D. Mureşan, A. Hangan, Road quality evaluation with mobile devices, *Proc. 2015 16th International Carpathian Control Conference (ICCC)*, Bucuresti, Romania, 2015, pp. 458–464. ⇒114

[10] L. Văcariu, A. Hangan, O. Creţ, A. Creţu, A decision support system on the cyberwater platform, *Proc. 2017 21st International Conference on Control Systems and Computer Science (CSCS '17)*, Bucureşti, Romania, 2017, pp. 591–598. ⇒ 112

[11] M. Xie, S. Han, B. Tian, S. Parvin, Anomaly detection in wireless sensor networks:a survey, *Journal of Network and Computer Applications* **34,** 4 (2011) 1302–1325. ⇒103, 104

[12] Y. Zhang, N. Meratnia, P. Havinga, Outlier detection techniques for wireless sensor networks: a survey, *IEEE Communications Surveys and Tutorials* **12,** 2 (2010) 159–170. ⇒103, 104

[13] ∗ ∗ ∗ Rompetrol Refinery explosion, *www.wall-street.ro*, http://www.wall-street.ro/articol/Companii/205623/rompetrol-rafinare-trimisa-in-judecata-in-dosarul-privind-explozia-de-la-petromidia-soldata-cu-doi-morti-si-doi-raniti.html. ⇒102

Received: September 7, 2017 • Revised: November 17, 2017