Anomaly Monitoring Framework Based on Intelligent Data Analysis

Prapa Rattadilok[1], Andrei Petrovski[1], and Sergei Petrovski[2]

[1] School of Computing Science and Digital Media, Robert Gordon University, UK
{p.rattadilok,a.petrovski}@rgu.ac.uk
2 School of Electric Stations, Samara State Technical University, Russia
petrovski@rambler.ru

Abstract. Real-time data processing has become an increasingly important challenge as the need for faster analysis of big data widely manifests itself. In this research, several Computational Intelligence methods have been applied for identifying possible anomalies in two real world sensor-based datasets. By achieving similar results to those of well respected methods, the proposed framework shows a promising potential for anomaly detection and its lightweight, real-time features make it applicable to a range of in-situ data analysis scenarios.

Keywords: intelligent data analysis, automated fault detection, big data, real-time, K-Means.

1 Introduction

Due to the fast pace of computing technology advancements in terms of both memory capacity and processing speed, increasing amounts of high precision data is becoming available. With this vast amount of data coming on stream, traditional data acquisition and data processing methods have become inefficient or sometimes inappropriate. Heterogeneous data sources also add complexity in the form of analytical challenges, especially when there exists time and/or cost differences in processing data from different sources. It has become more important than ever before (especially in a real time environment, where the processed information is only useful until a particular point in time and excessive latency would render the information useless [1]) to be able to effectively distil the large amount of data into meaningful information, as well as optimally selecting the data sources to minimise the time and/or cost.

The detection of outliers (or anomaly) is the process of finding patterns in a given data set that do not conform to an expectance and is the subject of much recent research [8]. This can be especially important in the case of big data where data volumes and sample rates limit the amount of data that can be simultaneously processed. Machine learning techniques are commonly used, but rely heavily on human knowledge integration as well as human involvement in defining every possible anomalous pattern, which are often unknown *a priori*. The resulting algorithms tend to be very

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 134–141, 2013.

problem-specific, and in most cases they are even designed specifically to evaluate pre-defined types of input, which makes them inappropriate in dynamically changing environment or in new problem domains.

2 Automated Anomaly Detection

Considering the example of automotive process control, various sensors may be used for different types of analysis. An anomaly in the air conditioning system, whilst the fan sensor indicates fully functioning blades, could imply a blockage, which requires manual/visual inspection inside the ventilation holes. Activating sensors in an ad-hoc basis and timely manner can help to reduce processing cost, as well as the amount of data available for analysis. Ideally, the data that is obtained from simple sensors should be chosen for anomaly detection. More detail-rich data from sophisticated sensors can later be activated, the resultant data undergoing further investigations. Fig. 1 illustrates the system overview which exploits both fast and detail-rich data.

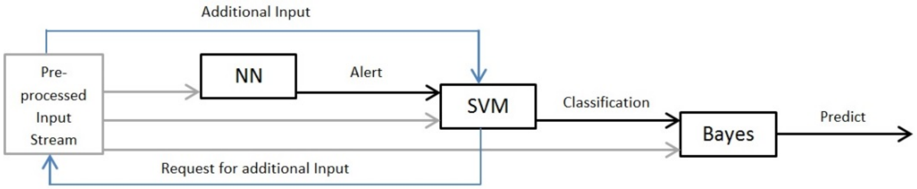


Fig. 1. System overview

The three Computational Intelligence (CI) techniques shown in Fig. 1, which are responsible for the identification, classification and prediction of anomalies, are Artificial Neural Network (ANN), Support Vector Machine (SVM) and Bayesian network (Bayes) respectively. This framework has been successfully applied [2], where the ANN operates on a fast data input stream. Experimental conditions (ie. learning rates, momentum and number of hidden units) were auto-configured using the standard WEKA configuration framework. Depending on the characteristics of the identified anomalies, either additional fast data may be required or detail-rich data may be processed to improve the accuracy of anomaly classification using an SVM, whereas the probabilities of future anomalies can be estimated using a Bayesian network. The abovementioned techniques are compared in [3].

3 Anomaly Identification Framework

To define properties and/or patterns of anomalies in advance is not a trivial task, especially within novel problem domains. It is also possible that what once was anomalous may become acceptable in a dynamic environment. One possible solution to pre-defining properties and/or patterns of anomalies within novel problem domains and/or dynamic environment is by combining the use of mathematical deviation analysis and computational intelligence techniques to evaluate the data. The mathematical deviation analysis allows the identification of undefined anomalies by evaluating the deviation from expectation, whereas computational intelligence techniques identify anomalies based on pre-defined patterns or provide continuous learning while solving the problem. Fig. 2 illustrates the anomaly identification process for the first sub-process (labelled NN) of the three-stage process in the system overview shown in Fig. 1.

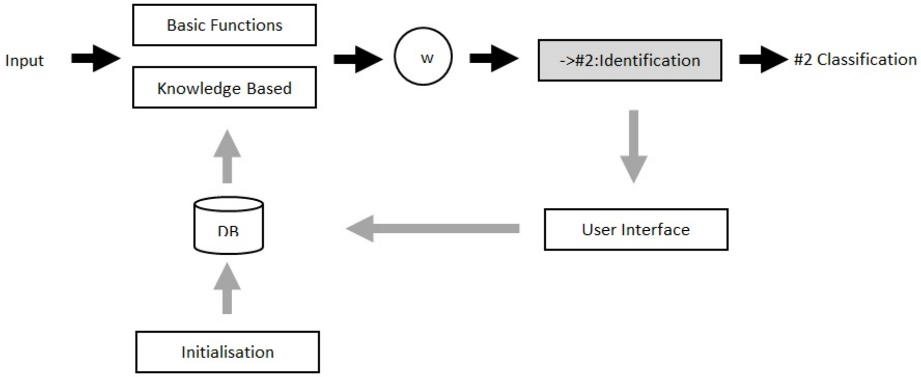


Fig. 2. Anomaly identification process

The "Knowledge-Based" process identifies an anomaly based on pre-defined patterns. Also, this process adaptively selects a suitable computational intelligence technique together with required parameters for the current problem solving stage. Any pre-defined patterns can be specified via the "Initialisation" process. The "Basic Functions" process mathematically evaluates the incoming input data in order to identify anomalies based on their deviation from the expectations.

Non-numerical input streams from sensors can all be manipulated into some form of numerical data. For example, images in video sequence can be evaluated in the sense of colour deviations or colour clustering, object trajectory, or even the intermittency of some input streams. The deviation analysis of numerical data allows for a more general approach to identifying anomalies, especially when the features of the anomalies are not known a priori.

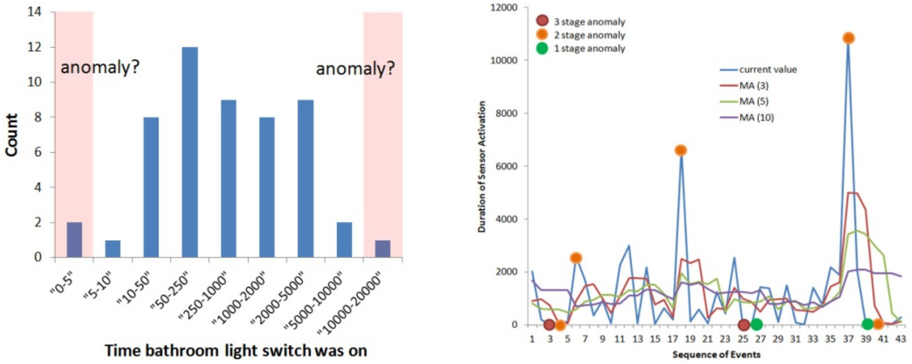
The results from both the "Basic Functions" and "Knowledge Based" processes are combined based on the degree of belief ("w" in Fig. 2) that varies over time depending on how accurate the "Knowledge Based" process is at identifying the anomalies. The "User Interface" process allows the user to provide anomaly confirmation/rejection responses. Only fast and/or explicit data will be used as the input for this anomaly detection process. The identified anomalies, as an output of the identification process, are passed onto the classification process, where association of anomalies with meaningful inferences can eventually be used to predict future process states. Additional inputs might be acquired at later stages to enhance accuracy.

4 Experimental Setup and Results

Two datasets from two different fields are used in the experiments: a multi-sensor smart home environment and an automotive process control. The application programs are developed using Java with the Encog library [4]. The mathematical deviation analysis for the “Basic Functions” process is evaluated using three different window lengths: short-, medium- and long-term. A number of computational intelligence techniques This included an ANN approach using multi-layer perceptron (MLP) with back propagation of error (back-prop) and an SVM approach using a sequential minimal optimization algorithm for training a support vector classifier; all experimental setup parameters are determined empirically.

4.1 Multi-sensors Smart Home Environment

The research of smart home environments has grown in recent years due to the aging population over the world and an increase in availability of inexpensive sensors. The motivation of the research is to ensure a safe living environment and lifestyle for smart-home occupants by monitoring the occupants’ movement, behaviour and interactions with objects/appliances on a frequent basis. The dataset used in this paper was obtained by non-invasive monitoring of object and appliance manipulation [5]. Changes in activity patterns, deviations in terms of regularity or duration of different activities are examples of possible anomalies, but without ground truth it is impossible to know for sure what should be classed as anomalies. Fig. 3a highlights possible anomalies at high and low values of the available sensors based on the duration, as well as regularity, of interactions. Fig. 3b highlights the anomalies detected using the ”Basic functions” process based on shorter-, medium, and longer-term memories of 3, 5 and 10 recent events respectively. The sensor output is considered anomalous if the current value is three times bigger or fifty times smaller than expected.



(a) **(b)**

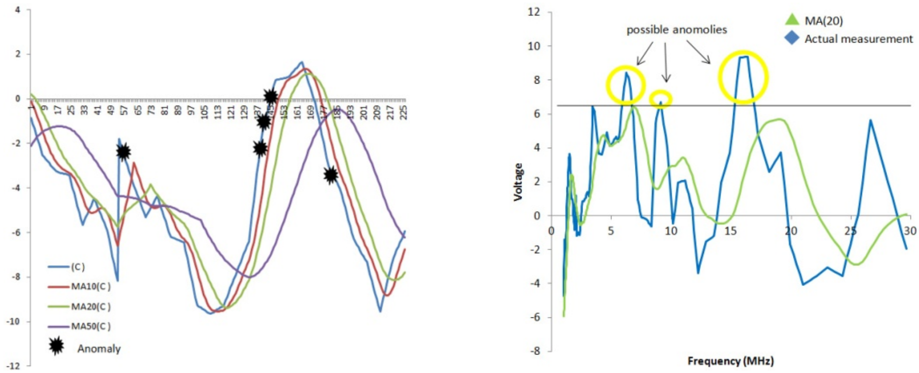
Fig. 3. Possible anomalies within the smart home dataset

Considering two examples: 1) at 9pm, on a Monday, the bathroom light was switched on for 5 seconds, and only bathroom cupboard sensors was activated (i.e. open and close); 2) at 11am, on a Monday, the bathroom light was switched on for 20,000 seconds, no other sensors are activated.

Information from other sensors and/or the context based information (e.g. time of day or day of week) can help with the identification process. The idea is to quickly identify possible anomalies based on the deviation analysis, and to exploit further the data related to the identified anomalies, if required, at the later stages. Fig. 3b illustrates that the best coverage of anomalies is obtained using the deviations based on the results of two out of three stages. When only one or all of the stages are used, the number of detected anomalies deteriorates. As previously stated, the anomaly confirmation/rejection depends on the user or expert knowledge. The "Knowledge Based" process is trained based on the anomaly confirmation-rejection responses from the user. The neural network accepts two inputs (i.e. current values and deviations) and produces the anomalous identification as an output. It is possible to achieve the accuracy rate of about 78% on an unseen test set using 5-fold cross-validation.

4.2 Automotive Process Control

The next dataset is based on the effect of an interference-suppression capacitor in terms of noise at different frequencies when the capacitor is connected to the bonding or to the engine cylinders. Anomalies in interference voltage can be detected in any of the three options of connecting the interference-suppression capacitor when the noise level is changing too abruptly (Fig. 4a) or significantly exceeds a threshold (Fig. 4b). The same "Basic Functions" process as applied to the multi-sensor smart home environment dataset is adopted in these experiments.



(a) (b)

Fig. 4. Possible anomalous interference voltage

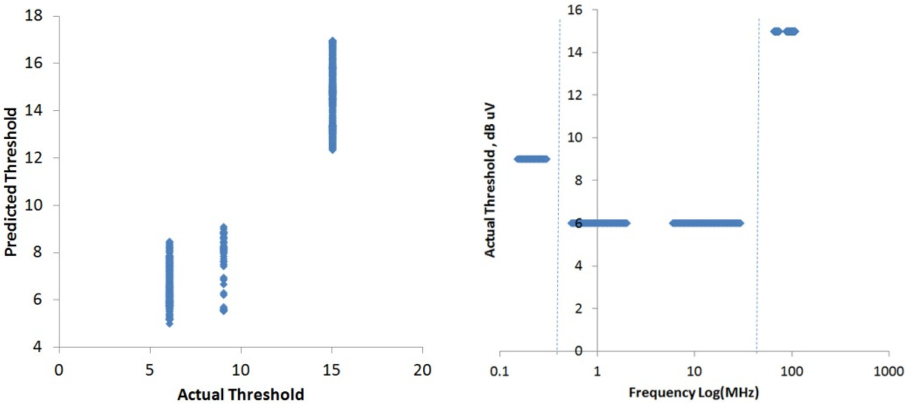
In Fig. 4a, the small, medium and large diapasons of frequency values are 10, 20 and 50 MHz respectively. Instead of using a single value, the intervals of ±1.25, ±2.5 and ±7.5 dBμV are used to compare between the actual value and the mathematical deviation. If the differences are larger or smaller than the specified limits in at least two of the three ranges of frequencies, then the value is considered anomalous. In Fig. 4b, anomalies are determined by comparing the actual values to the highest peak obtained using the deviation analysis.

Fig. 4a is derived using the data on the interference voltage for frequencies above 65 MHz when interference-suppression capacitor connected to the engine is used. As shown, the anomalous frequency diapasons are 53-56, 133-150 and 177-180, this equates to the frequency ranges of 86.97-87.30, 96.25-98.34 and 101.76-102.15 MHz. Fig. 4b is derived using the data on the interference voltage for frequencies below 30 MHz when no interference-suppression capacitor is used. As can be seen from the figure, three anomalies are identified. The result coincides with the expert knowledge obtained related to excessive noise as discussed in Fig. 5a and 5b.

In the dataset used, the threshold values for acceptable noise are not specified for all the. Fig. 5a and 5b demonstrate the results of applying two computational intelligence techniques to estimating the missing thresholds. Fig. 5a compares the actual threshold and the predicted threshold using the "Knowledge Based" process together with a linear regression technique (Eq. 1), where the variables $V_i, i = \overline{1,4}$ are frequency, interference voltage with suppression capacitor connected to the bonding, interference voltage with no suppression capacitor, and interference voltage with suppression capacitor connected to the engine cylinders respectively.

$$T = (0.0825 * V_1) + (0.0634 * V_2) + (-0.1852 * V_3) + (0.3486 * V_4) + 9.7531 \qquad (1)$$

Fig. 5b represents the thresholds of 6, 9 and 15 as semantic classification targets rather than linear values. The value of 47.3 is the mid-point between the lowest value for any threshold 15 sample (65) and the highest value for any threshold 6 sample (29.6). The value of 0.416 is the mid-point between the lowest value for any threshold 6 sample (0.5333) and the highest value for any threshold 9 sample (0.2982).



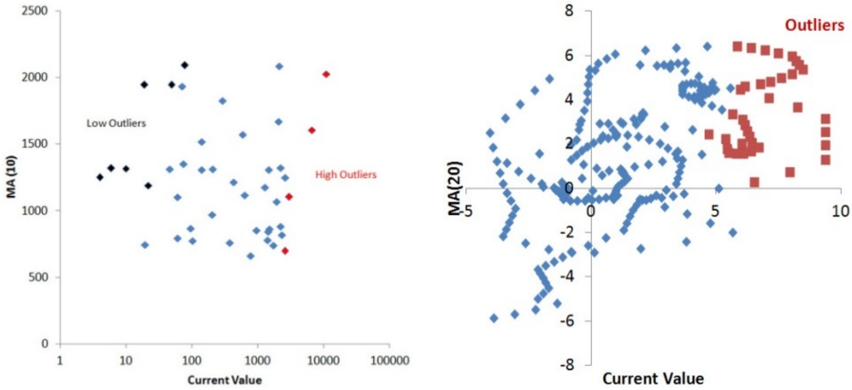
(a) (b)

Fig. 5. Using computational intelligence techniques to determine interference voltage threshold

From Fig. 5a, it is apparent that samples with a threshold of 15 are accurately predicted, but there are some (60) mis-classifications of samples with a threshold of 6 and 9. This suggests some non-linearity in the data that is difficult to predict using linear regression. For 'Big' data this initial approach of building a linear regression model is useful because it enables assessing difficulty and non-linearity of the relationships within the dataset. A non-linear regression approach may be more accurate. Fig. 5b confirms the finding when frequency is plotted against the actual threshold. The clear segregation of the three thresholds based of frequency is apparent. There also appears to be at least 2 sub populations for the samples with the threshold of 6 and 15.

4.3 K-Means Comparisons

Due to the lack of ground truth about anomalies in the datasets, algorithmic validation and performance comparison are achieved through assessing the level of agreement with other methods. Clustering is a well-established method for detecting outliers and anomalies in a dataset [6]; a constrained k-means version [7] of the clustering algorithm is applied to the data. The results of the clustering approach on the smart home environment data and automotive process control data are shown in Fig. 6a and 6b respectively. Similar results are achieved by clustering as by the proposed approach.



(a) (b)

Fig. 6. Comparisons with K-means clustering

In Fig. 6a, by specifying three clusters, high and low outlier regions to the core data can be generated. These two outlier clusters contained all the values highlighted in Fig. 3b plus one extra high outlier and one extra low outlier. In Fig. 6b, two clusters are specified, in order to identify the values above and below the threshold. The outlier clusters contained all the values highlighted in Fig. 4b. So while the results of the two methods are similar, the proposed approach has several advantages that include the ability to detect anomalies in real time without the need to carry out clustering on the whole dataset and the comparatively lower computation load. Both of these advantages are especially important for real time monitoring scenarios.

5 Conclusions

Activating various sensors and processing their data incurs different costs. With the large volumes of data from heterogeneous sensors, evaluating explicit sensor data in order to identify an anomaly before exploiting detail rich and/or more expensive sensor data can be beneficial, especially when (near) real-time decisions are required.

In many cases, a specific problem domain can be unexplored, and the generality of the approach can boost the applicability of the system with little or no modification. The deviation analysis is combined with computational intelligence techniques to tackle both explored and unexplored problem domains. The approach has been successfully applied to two datasets from very different fields of knowledge.

When comparing the proposed approach with a K-means clustering technique, comparable results are obtained. This supports the validity of the approach in identifying anomalies by using the proposed deviation analysis. Furthermore, the approach is generic enough to tackle different datasets from two very different problem domains.

Currently, all parameters are determined empirically both in the "Basic Functions", as well as the "Knowledge Based", processes. Various sensors may require different settings, and these values may vary over time. The acceptable range of deviation may change over time and between various sensors. Automatically determining optimal parameters, as well as adaptively adjusting them in real time, would improve the versatility of the algorithm with little or no modifications. This will be investigated further, and the identification output will be added to the next two sub-processes, classification and prediction, providing better insight into the nature of various anomalies.

References

1. Jacobs, A.: The Pathologies of Big Data. ACMQueue (2009)
2. Rattadilok, P., Petrovski, A.: Inferential Measurements for Situation Awareness: Enhancing traffic Surveillance by Machine Learning. In: CIVEMSA 2013 (to appear, 2013)
3. Correa, M., Bielza, C., Pamies-Teixeira, J.: Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Systems with Applications 36(3), 7270–7279 (2009)
4. http://www.heatonresearch.com/encog
5. http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/
6. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85–126 (2004)
7. Tung, A.K.H., Han, J., Lakshmanan, L.V.S., Ng, R.T.: Constraint-Based Clustering in Large Databases. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 405–419. Springer, Heidelberg (2000)
8. Dereszynski, E.W., Dietterich, T.G.: Probabilistic models for anomaly detection in remote sensor data streams. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence. arXiv:1206.5250 (2012)