A taxonomy and platform for anomaly detection

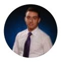
Conference Paper · May 2018

DOI: 10.1109/AQTR.2018.8402710

4 authors:

Gheorghe Sebestyen
Universitatea Tehnica Cluj-Napoca
63 PUBLICATIONS 229 CITATIONS
SEE PROFILE

Anca Hangan
Universitatea Tehnica Cluj-Napoca
35 PUBLICATIONS 117 CITATIONS
SEE PROFILE



Zoltan Czako
Universitatea Tehnica Cluj-Napoca
3 PUBLICATIONS 1 CITATION
SEE PROFILE

György Kovács
Universitatea Tehnica Cluj-Napoca
3 PUBLICATIONS 1 CITATION
SEE PROFILE

A Taxonomy and Platform for Anomaly Detection

Gheorghe Sebestyen, Anca Hangan, Zoltán Czakó, György Kovács
Department of Computer Science
Technical University of Cluj-Napoca
Romania
Email: Gheorghe.Sebestyen@cs.utcluj.ro, Anca.Hangan@cs.utcluj.ro

***Abstract*—There are hundreds of anomaly detection methods developed for different purposes and using a wide range of theoretical backgrounds, from system theory and signal processing towards artificial intelligence techniques. Therefore, it is very difficult for a specialist in a given domain (e.g. finance, industrial engineering, networking, environment monitoring, etc.) to select an anomaly detection method that fits best for a given application. The goal of this paper is to propose a taxonomy for anomaly detection methods and also to present a platform that allows a developer to find and tune a given anomaly detection method that is optimal for an application. The platform provides the basic functionalities needed to acquire, process and visualize multidimensional data collected from different sources, as well as the means to test and compare different anomaly detection techniques.**

***Index Terms*—anomaly detection, outlier detection, tools for anomaly detection, taxonomy**

I. Introduction

In the last decades, as more and more parts of human activity are taken over by automated and computerized systems, the task of detecting abnormal system behaviors should be also automated. An automated system (e.g. an energy supply system, an industrial process or a computer controlled social-financial system) should identify and react to such abnormal behaviors also in an automated way. But, the task of recognizing an abnormal behavior is not trivial even for a human observer. It takes a lot of experience, technical knowledge and intuition to make the difference between normal and abnormal behavior. In an automated anomaly detection system all these elements should be formulated as a coherent and efficient detection system.

The complexity of the anomaly detection process is given by a number of factors: it is difficult to find a unique discriminant for normal and abnormal behavior, the analyzed process data is complex (reflects a multi-state and multi-dimensional system) and in many cases the shape/pattern of the anomaly is a-priori unknown. Therefore, today there are hundreds of anomaly detection methods but neither of them is perfect. Some of them are proper for probabilistic and statistical data, others are better for time or space related data. In each case the detection method relies on some kind of correlation and redundancy present in the acquired data. Sometimes this is a statistical correlation, a spatial-temporal correlation or a functional correlation between a process parameters.

Many survey articles [2], [3], [4], [5], [6] tried to classify and in some way to introduce a given order between existing anomaly detection methods. But most of these surveys concentrate their search in a given application domain (e.g. economical data, networking, industry data, etc.) or emphasize methods based on a given theoretical approach (statistics, signal processing, artificial intelligence). Our goal is to unify all these efforts in a coherent anomaly detection taxonomy. Our goal is not to include all possible methods that exist today in the scientific literature, but to define some general principles that allows a multi-criteria classification of the existing methods.

As a practical result of such a classification attempt is the development of an open platform [7] that incorporates examples of anomaly detection methods from all the defined classes. The platform could be used as a primary tool to verify which anomaly detection approach fits best for a given system of set of data. Through its functionalities the platform gives a developer the possibility to build-up a sequence of steps for anomaly detection and allows him/her to evaluate the results form qualitative and execution efficiency perspective. So the goal of this paper is twofold: to propose a taxonomy of anomaly detection methods and to present the architecture and the main functionalities of an anomaly detection platform.

The rest of the paper is organized as follows. Section II includes related work. The proposed taxonomy is presented in section III. The architecture of the platform for anomaly detection is described in section IV. Section V includes the experiments, while section VI concludes the paper.

II. Related work

Anomaly is a general term that may cover a multitude of causes that bring a system out of its normal functioning conditions [5], [6]. Due to anomalies, monitoring data sets will include outliers. The term ”outlier” was originally used in the field of statistics and is defined in [1] as an observation that is inconsistent with the set of data it belongs to. Even though outlier values in datasets are the effect of anomalies in the systems functioning, in many cases the terms ”outlier detection” and ”anomaly detection” are used with the same meaning. There is a general consensus that anomaly detection is a vast topic that has applications in various domains and that there is no general solution to the anomaly detection problem. The solutions vary with the characteristics and particularities of each problem. In this context, there has been a lot of interest

978-1-5386-2205-6/18/$31.00 ©2018 IEEE

towards classifying anomalies and detection methods. Even so, many classifications take into consideration only specific application domains or a certain types of problems. A good example that supports this observation is the approach presented in [3], where the authors make a classification of anomaly detection problems and methods used in wired networks. Moreover they propose a general method for anomaly detection in wired networks that is based on pattern recognition. They construct a representation of normal behavior to be used as the pattern of correctness to identify abnormal behavior.

The idea of treating anomaly detection as a pattern recognition problem is taken further in [2], where the authors have a more general approach. Anomalies are defined as ”patterns in data that do not conform to a well defined notion of normal behavior” and anomaly detection is viewed as the problem of identifying those patterns. The authors propose a classification of anomaly detection solutions starting from specific application domains. They argue that the selection of a specific anomaly detection technique should be based on anomaly type, the nature of data and on the existence of labeled data. Finally, using this approach, they identify the most appropriate detection techniques for each application type in their review. The authors of [2] have a more general approach on the process of solving an anomaly detection problem by introducing general anomaly types and by identifying some general characteristics of datasets. However, rather than classifying the anomaly detection methods, they try to give a solution on how to select the methods for a specific application domain.

A classification of anomalies and a taxonomy for anomaly detection methods in wireless sensor networks is presented in [10]. The classification takes into consideration many characteristics of an anomaly such as source, type of anomaly, input data type, which are specific for sensor networks. Using a more general approach, the methods are classified based from the perspective of the field of computing they belong to: statistics, neural networks and machine learning.

At the opposite pole, there are classifications that are too general. For example, in [12] the authors make a general classification of outlier detection methods in which they take into consideration only the characteristics of the dataset. They identify three main classes of methods, in which there is no previous knowledge about data (unsupervised), in which there are models only for normality (semi-supervised) and in which there are models for both normal and abnormal behavior (supervised). Furthermore, they make a review of the methods used for outlier detection that belong tho the three classes they have identified, without discussing the variations that may appear when taking into consideration the type of anomaly or the source of the outlier value.

III. Taxonomy of anomaly detection techniques

In our opinion, most of the classifications of anomaly detection techniques present in a number of previous surveys, give an unilateral view of the existing techniques, a view based on a single criterion. This criterion was the application domain, the basic technique used for detection, the nature of the anomaly or the kind of analyzed data. Through this paper we promote a multi-view classification approach in which a given detection method is classified (and identified) through a number of equally important criteria. This taxonomy allows a developer to filter and finally to identify the anomaly detection method which best fit with its search interests. Than using the proposed platform the selected method(s) can be tuned in order to get the best detection results in the shortest possible time.

So, let’s analyze which are the criteria that can define a given anomaly detection method. Based on the existing literature we identified the following criteria:

A. The nature of the anomaly

Using this criterion we can identify the following classes:

- one point anomaly - a point (collection of attributes of a system at a given time) that is very different from the points considered normal; statistical and normal distribution techniques are proper for this kind of anomaly detection; here there are 2 subclasses:
  - some techniques consider the relation between a point and all the other points in the set;
  - the others consider the relation (e.g. distance) between a single point and the members of previously detected classes of points; here clustering techniques together with different distance metrics are combined for detection
- contextual anomaly - a more complex anomaly that considers the relation between the analyzed point and its neighbors; the techniques used here takes into consideration seasonal changes in the dataset; a special task here is to determine automatically or in an iterative way the optimal length of the window that defines the context of a point
- sequential data anomaly - considers not the value of an individual point but the multidimensional shape of the point sequences ordered in time; typical example here is to recognize abnormal sequences in ECG signals; the techniques used here may be built upon features extracted from signal analysis (e.g. FFT, correlation, etc.) or methods that can detect similarities and disparities in string sequences

B. The nature of the analyzed data

In this classification we can use a number of sub-criteria:

- Number of attributes characterizing a data-point
  - Univariate points - points; characterized by a single parameter value
  - Multivariate points - a point characterized by a number of attributes or parameter values; the attributes may be numerical, textual and mixed.
- Correlation between data-points
  - Statistical correlation - data points that follow some statistical or probabilistic rules (see stochastic signals);

- time correlation - data represent time series where a data-point is correlated with the previous samples of the same parameter; filtering, moving average or auto-regression techniques may be applied for anomaly detection; there are also techniques that try to identify the model of the system generating previous data-points and compare the estimated value (computed with the model) with the actual value; an anomaly is detected when the difference exceed a predefined threshold;
- spatial correlation - is similar with the previous case with the difference that the correlation is between spatially neighboring points; this case is typical for data collected from sensor networks where a physical connection exist between neighboring sensors;
- functional correlation - in this case there is a functional relation between different parameters measured in a system (e.g. current, voltage and power consumption in an electrical distribution network); the correlation can be derived from the physical laws governing the process or it can be deducted with inter-correlation computations; if a correlation exist between two parameters it can be exploited for anomaly detection; in case of causal correlations a ”drastic change in the ”effect” signal is not considered an anomaly if it’s ”cause” signal change it’s value as well.

C. The nature of the detection principles

Anomaly detection methods may be developed upon a variety of generic techniques borrowed from different research/science areas: statistics, artificial intelligence, data mining, pattern recognition, system theory, signal processing and many others.Below we give some potential techniques which were successfully adopted for anomaly detection:

- Statistical methods - consider that the collected data should have a given probability distribution; an outlier value (abnormal data) is out of a range considered normal (e.g. mean value +/- 2* standard deviation); more complex methods improve the accuracy by considering seasonal variations in the dataset; these methods are typical for financial data collections;
- Artificial intelligence methods - include different search, classification, clustering and data mining techniques grouped under the generic name of artificial intelligence; typical for this category of detection methods is that the rules for discriminating between normal and abnormal values are not given; the algorithm tries to identify some rules (discriminants) for outlier recognition; depending on the learning method there are two approaches:
  - Learning method
    - supervised techniques - the rules or discriminants are learned from positive and negative examples (data samples are labeled as normal or outlier);
    - semi-supervised techniques - only the normal behavior can be learned from positive examples;
    - unsupervised techniques - the learning process cannot consider some a-priori specified anomalies.
  - Methods applied
    - clustering and classification - the detection method considers that an outlier is not ”close” enough to the defined or determined data clusters; there is a variety of methods that consider different metrics for determining the proximity of a data to a given cluster (e.g. distance or local density measure)
    - shape and pattern recognition - here, not the individual data but the shape of the signal (sequence of samples) makes the difference between a normal or an abnormal state; SVM (support vector machine), neural networks or Markov chains are trained in order to make the difference between normal and abnormal shape
    - string matching - in this case the evolution of a signal is recoded with a finite number of symbols (e.g. characters) and the problem of identifying an anomaly is reduced to the problem of detecting a string that does not fit with the strings considered normal
- System theory and signal processing methods - these methods [6], [9] work with time series or with spatial-temporal data; some of the methods use more or less simple filtering techniques in order to detect outliers; other methods try to model the behavior of the system generating the data and compare the predicted value (based on the model) with the actually provided data
  - min-max, threshold - these are the simplest forms of anomaly detection, where minimal, maximal or threshold values are determined from a sequence considered normal; values exceeding these limits are considered anomalies;
  - filtering - by comparing a signal with its low-pass-filtered version may give an indication of an outlier value; the idea behind the method is that an outlier value (e.g. a noise in the signal) does not preserve the ”continuity” of the signal;
  - modeling - here, using system identification techniques a model of the process generating the data is built and than the model is used to predict the the next values; if there is a significant error between the collected data and the predicted one the conclusion is an outlier data; there are different techniques used for this purpose like: autoregression, autocorrelation, multivariate correlation, etc.

D. The nature of the application domain

Sometimes the type of anomaly detection method used in a given case is determined (at least partially) by the domain of the application. For instance in case of financial data the statistical methods are recommended or in case of industrial applications time series related methods are used. Therefore, one classification criterion for anomaly detection methods is the application domain. Here are some of the domains where anomaly detection play an important role: economical and financial data; cyber-physical systems and industrial processes [9]; environmental data; weather forecast; geo-spatial data; medical diagnoses; medical diagnoses; computing and networking (virus detection, intruder or malicious code detection).

Depending on the application domain the quality criteria used for selecting the best detection method may differ; in some domains few false positives are not critical (e.g. weather forecast, environmental data), but in others precision is essential (e.g. medical domain, industrial processes). In some domains the detection must be done on-line (e.g. cyber-physical systems, industrial processes), restricting the search window and the available execution time; in other cases off-line processing is done with no critical time restrictions (e.g. financial data, medical diagnoses), allowing a more accurate and consequently more time consuming detection method implementation.

E. Final remarks regarding a taxonomy of anomaly detection methods

As the above presentation shows, the multitude of anomaly detection methods present in the scientific literature cannot be classified using a single criterion or viewpoint. Multiple criteria should be applied in order to position a given method on the multidimensional anomaly detection ”map”. In our view the dimensions of this map or taxonomy are as follows:

- the anomaly type addressed by a given method (e.g. singular, contextual, shape-related, etc.)
- the kind of data under analysis (e.g. statistical data, time series, spatio-temporal data, etc.)
- the theoretical background of the method (e.g. system theory, statistics, artificial intelligence, string theory, etc.)
- the application domain (e.g. economics, industrial, environmental, medical, etc.)

In every major dimension there are subdivisions as specified in this chapter. With this map in mind it is much easier for a developer to address a given practical anomaly detection problem, select the best choices and make efficiency comparisons. This taxonomy helped us to build a coherent platform for anomaly detection.

IV. A General Architecture of the Platform for Anomaly Detection

The Anomaly Detection Platform (ADP) developed by our team is meant to offer a pragmatic tool for specialists involved in some kind of anomaly detection process, a tool meant not as a final solution but as a starting point in the process of finding the best method that fits the quality and efficiency criteria of a given application domain. Therefore the platform contains those basic functionalities that allows a specialist to collect, process and visualize data for anomaly detection purposes. Here are the functionalities we considered necessary for such a platform:

- Data harvesting tools - that allows acquisition of data from a variety of datasets having different formats (Excel, CSV, ARFF) or from different physical sources (sensor networks, smart devices, IoT); a special treatment is given to real-time data collection and anomaly detection methods;
- Data preprocessing tools - instruments meant to transform the raw data into a noise free and normalized data; typical methods included in this category are: parameterized filters, transforms (e.g. FFT, wavelet) or histograms; there are also methods for determine some statistical parameters of the input data such as: min-max, median value, standard deviation, etc.
- Anomaly detection methods - a wide group of methods that try to cover the most representative nodes of the presented taxonomy; our goal is to offer a developer a rich set of possibilities from were to chose and compare; the open nature of the platform allows new methods to be added to the existing library;
- Visualization tools - very important in the process of finding the best anomaly detection method because they offer a bi-dimensional representation of otherwise multidimensional data, much easier to understand for the human eye;
- Generators of datasets and artificial signals - necessary in the process of validating or measuring the quality of some new detection methods; special techniques are applied in order to combine ”normal” data with artificially created anomalies.

The tools mentioned above are integrated into an open architecture platform allowing continuous extension with new methods.A unique internal predefined data format assure interoperability and interchangeability between the existing tools and functionalities.

According to the platform’s ”philosophy” an anomaly detection scenario can be built (in an interactive way) as a chain of procedures, where the output of a procedure becomes the input for the next stage.The user, through the platform’s interface, can combine multiple procedures in a single scenario.

As part of the platform we included multiple datasets from different application areas that are offered as benchmarks by groups of researchers working in the anomaly detection field. These datasets offer the variability necessary for testing a given new detection method from different perspectives and in comparison with some existing methods (considered as baselines).

Fig.1 shows the application’s components (toolkit) for detecting anomalies that can be accessed through a user interface. An important concern for us was to find different way of representing the input data as well as the final result (e.g. anomaly points marked in a detectable way) that allows a human observer to identify the discriminant characteristics that make the difference between the normal and the outlier data points. In our view if the human eye can detect a difference than there is a higher probability of defining a formal method that can detect the same difference. But in case of multivariate

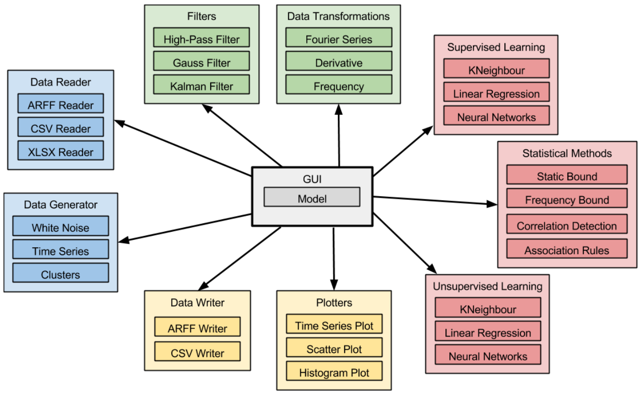


Fig. 1. Components of the anomaly detection platform.

datasets this task is not trivial. The anomaly cannot be detected as a significant (visible) change in the characteristics of a given attribute (or small group of attributes), it is a consequence of a combination of many attributes (which is harder to observe on a 2D or 3D representation). Techniques like "principal component analysis" or Adaboost are needed to select or emphasize those attributes that can discriminate between normal and outlier data.

Another issue that guided the development of the platform was to offer the user the possibility to change the parameters of a given method (e.g. length of a filter, number of clusters, etc.) and to compare the results in order to find and tune the most efficient method for a given dataset. Even if there are attempts to find a universal method good for multiple cases, our experience showed that different application areas and consequent datasets may require particular solutions. This observation is more important if we consider that the best solution for a given application should be a compromise between detection quality (e.g. precision, recall) and execution time (in most cases two contradictory criteria). This is more critical in case of real-time detection requirements or when the available computing resources available for detection are limited (e.g. control applications implemented on micro-controllers).

V. Experimental results

The purpose of this section is to show through some examples the usage possibilities of the anomaly detection platform. Here we emphasize three aspects: the ability to visualize in a relevant way multivariate datasets, the possibility to chose between different anomaly detection techniques and the ability to tune the parameters of a given method in order to obtain the best results. Fig. 2 shows the 2D representation of a number of multivariate datasets (proposed in [11]) obtained using the "principal component analysis" (PCA) method. As it can be seen the normal (blue dots) and the abnormal (red dots) data points are well separated in each of these views. The PCA method was used for visualization purposes as well as a preprocessing step for a faster anomaly detection process. The next experiments were made on the same dataset that reflects information related to breast cancer [11].This dataset

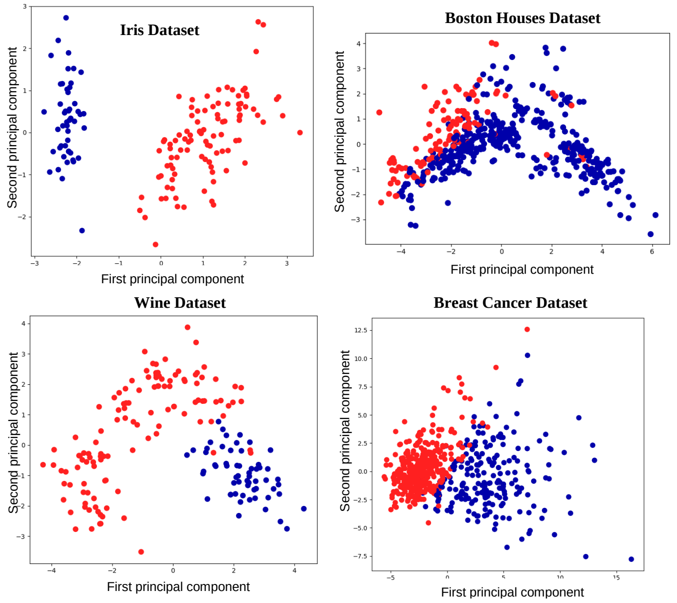


Fig. 2. PCA results for different datasets.

contains 569 data points and each data point has 32 attributes. The dataset was labeled showing normal (e.g. benign) and abnormal (e.g. malign) data points. A part of the dataset was used in the training (learning) phase and the rest in the testing phase. The graphic representations presented in fig. 3 show the obtained qualitative results for different detection methods. The x axes represent one variable parameter of a given method (e.g. gamma for the SVM or alpha for the NN, see below) and the Y axes represent the accuracy of the detection measured with equation below:

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} eq(y, \hat{y})$$

where n is the number of samples, y is the actual label (normal or abnormal), ŷ the estimated label and eq(y,ŷ) is 1 if y=ŷ and 0 otherwise.

In order to find the best anomaly detection method 4 generic methods were tested:

- kNN (k nearest neighbors) with different distance metrics
- NN (neural networks) with 2 hidden layers and 10 nodes
- RF (random forest) with variable depth
- SVM (support vector machine) with variable C and gamma parameters

Fig. 3 shows the best results (best configuration) obtained with the 4 methods. This comparison allows a developer to chose between multiple generic choices. In our case for the given dataset a number of methods if well tuned give relatively high accuracy values (over 0.95).

The next set of experiments were meant to tune the configurable parameters of a given generic method. Fig. 4 shows the qualitative results obtained with the SVM method for different values of the C (node importance) and gamma parameter. Based on these diagrams the developer can decide for the best

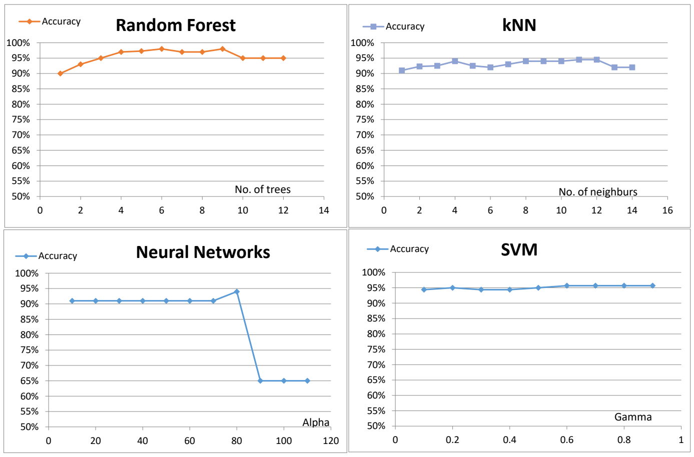


Fig. 3. Best results obtained with RF, kNN, NN and SVM

combination of parameters that give a high accuracy on the test set and assures a good robustness (less quality variation in the neighbor of the selected values).

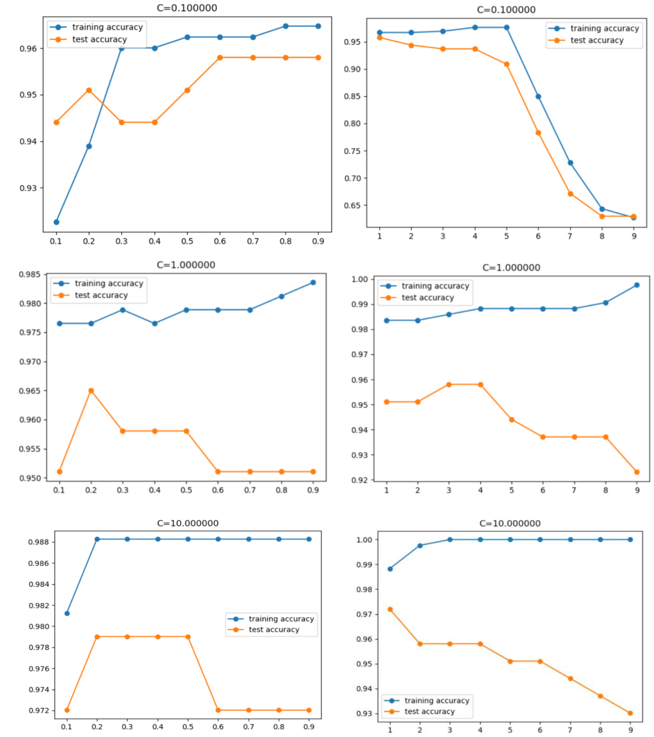


Fig. 4. Fine tuning of the SVM method (C=0.1, 1, 10; gamma between 0.1 and 10)

VI. Conclusions

The taxonomy proposed in this paper tries to organize the multitude of anomaly detection methods in a coherent way, allowing a developer to select the best methods that fit to the requirements of a given application. This taxonomy was used as a theoretical background for the development of a platform that incorporates different anomaly detection techniques. The platform implements the main functionalities needed by a developer in the process of finding the best strategy for a given domain that assures high quality anomaly detection. It includes multiple data acquisition, preprocessing, anomaly detection and visualization functionalities that may be combined in a specific processing flow. The platform has an open and scalable architecture allowing extensions for each of the mentioned groups of functionalities. Every individual functionality is implemented as an autonomous service and multiple services may be orchestrated in a logical flow in order to obtain the final anomaly detection results. The richest group of functionalities is that containing the anomaly detection methods. Here, we tried to cover most of the development directions present today in the scientific literature, from statistical methods, towards, signal processing and artificial intelligence. The platform can be used as a training tool for those who develop anomaly detection solutions. For the same learning purposes we included a multitude of datasets that cover a diversity of application domains. The experimental part of the paper demonstrates some of the functionalities of the platform, including the possibility to compare the result obtained with different methods.

Acknowledgment

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.

References

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

[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A Survey, ACM Com- puting Surveys 41, 3 (2009).

[3] J. M. Estevez-Tapiador, P. Garcia-Teodoro, J. E. Diaz-Verdejo, Anomaly detection methods in wired networks: a survey and taxonomy, Computer Communications, 15. October 2014, volume 27

[4] M.A. Rassam, A. Zainal, M.A. Maarof, Advancements of Data Anomaly Detection Research in Wireless Sensor Networks: A Survey and Open Issues. Sensors 2013, 13, 10087-10122.

[5] S. Agrawal, J. Agrawal, Survey on Anomaly Detection using Data Mining Techniques, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, ed. Elsevier,Procedia Computer Science 60 (2015) 708 713

[6] M. Gupta, J. Gao, C. C. Aggarwal,J. Han, Outlier Detection for Temporal Data: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2014

[7] N. Laptev, S. Amizadeh, I. Flint, Generic and Scalable Framework for Automated Time-series Anomaly Detection, KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 1939-1947, Sydney, 2015

[8] S. Cateni, V. Colla, M. Vannucci, Outlier Detection Methods for Industrial Applications, chapter in book "Advances in Robotics, Automation and Control", book edited by Jesus Aramburo and Antonio Ramirez Trevino, ISBN 978-953-7619-16-9, Published: October 1, 2008

[9] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, UCR Time Series Classification Archive,www.cs.ucr.edu/ eamonn/time-series-data/

[10] Y. Zhang, N. Meratnia, P. Havinga, Outlier Detection Techniques for Wireless Sensor Networks: A Survey, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, VOL. 12, NO. 2, SECOND QUARTER 2010

[11] M. Lichman(2013). UCI Machine Learning Repository.

[12] V.J. Hodge, J. Austin, A Survey of Outlier Detection Methodologies, Kluwer Academic Publishers, 2004