IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identification Of Outliers Based On Sensitivity Of Data

Prof. Shiva Sumanth Reddy | Amogh S Bharadwaj | Chandrakanth N Murthy | H Vishwajit | Harsh V Challa

Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore, India.

Abstract- Outlier detection is examined and can be applied in many domains. These Outliers occur because of human error while data entry, fraudulent error when a malicious practice occurs, system behaviour, by natural deviations in datasets or instrument error. The detection of outliers has been used over many years, to identify outliers and analyse outliers where required. Sometimes the outliers need to be removed, Sometimes we use it to identify Outliers. The main challenge in outlier detection is to work on data that is highly Sensitive data. Sometimes the data is so sensitive that the outlier data coincides with the normal data, this usually occurs in the domain of malicious activities. The proposed system assists to clean data in less time and great accuracy. Our paper focuses on outlier detection which can be applied to various domains that have time series data. This shows critical review on various approaches to detect outliers and provides the most accurate technique for particular type of data.

Index Terms- Anova Test, Box plot, Data Analytics, Dataset Sensitivity, Isolation Forest, Local Outlier Factor

1.INTRODUCTION

An Outlier is data entry that deviates drastically from the normal data. The outliers are based on their sensitivity, if the outliers coincide with normal data it is highly sensitive and if the outliers deviate from normal data it is moderately sensitive. Detecting outliers may be used in a broad range of fields, including system defect detection, intrusion detection, credit card fraud detection, insurance fraud detection, and monitoring for enemy activity in the military.

1.1 Types Of Outliers

Outliers are basically classified based on their occurrences namely Point Outlier, Contextual Outlier and Collective Outlier.

1.1.1 Point Outliers

When a single data instance that is sitting outside the normal range of the dataset. For instance, if a patient's height was recorded but one digit was left off, the dataset would include a point outlier.

1.1.2 Contextual Outliers

When the data instances are drastically different from the dataset, but only within a specific context. For example, in a dataset of Bangalore temperatures over time. A temperature reading less than fifteen degrees in winter is normal but the same dataset reading below fifteen in summer is an outlier.

1.1.3 Collective Outlier

When a series of data instances drastically drift from the normal form of the dataset.For Example, Using a time series to highlight seasonal or daily variations in subscribes and unsubscribes to an email marketing list. The level of subscribed users might be labelled as a collective outlier if it remained constant for several weeks without variation. It is common for individual users to unsubscribe and for new users to subscribe, therefore a static figure would be considered an oddity.

1.2 Types Of Datasets

1.2.1 Cross-sectional data

When several people are observed at the same time, cross-sectional data is obtained. Cross-sectional data may include observations made several times, although in such circumstances, time itself has little bearing on the study. An example of cross-sectional data is the test scores of students in a particular year. Another example of cross sectional data is the GDP of nations in a specific year. Another illustration of cross sectional data is data used for customer turnover analysis.It should be noted that the exam scores of students and the GDP of countries are cross sectional datasets because all observations were made within a single year. In both instances, the cross-sectional data essentially serves as a picture of the world at a particular moment in time. But data on customers for churn research may be gathered over longer periods of time, including years and months.Though time may not be a significant factor for analytic purposes, But customer turnover data may be gathered at several periods in time, it may still be regarded as a cross-sectional dataset.

The following graph, which uses the example of military spending as a proportion of GDP for 85 nations in 2010. We guarantee the data's cross-sectional character by using data from a single year. To illustrate various statistical characteristics of the military spending data.

The figure clearly shows that military spending has a large peak at around 1.0% and is slightly left-skewed. Near 6.0% and 8.0%, a few tinier peaks may also be seen.

Figure 1:Military Expenditure for 85 countries in 2010

1.2.2 Time-series data

A time-series data is a collection of information or observations made at either regular or irregular periods of time. Usually, time series is often a set of data points recorded at regularly spread out intervals. The count of data points that are captured may occur hourly, daily, weekly, monthly, quarterly, or yearly.

Applications of time series can be found in business, finance, or statistical fields. The daily closing value of stock indices like the NASDAQ or SENSEX is a widely popular example of time series data. Time series are also often used in econometrics, signal processing, pattern identification, sales and demand forecasting, weather forecasting, and earthquake prediction.

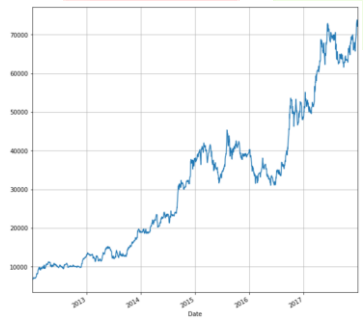

Figure 2:Senex value from 2012 to 2018

The given figure illustrates a time series dataset for change in the value of sensex for a fixed period of a year

1.2.2 Panel data

Data that is observed on various cross sections across time is known as panel data, often known as longitudinal data. It is an assortment of measurements acquired from many sources, accumulated over regular time periods, and arranged chronologically. Several examples of the types of groups that can be included in panel data series are countries, organisations, people, or demographic categories.

Similar to time series data, panel data also comprises observations that were sequentially acquired at predetermined intervals. Panel data involves observations made across a group of individuals, just like cross-sectional data does.

Person	Year	Income	Age	Sex
1	2013	20,000	23	F
1	2014	25,000	24	F
1	2015	27,500	25	F
2	2013	35,000	27	M
2	2014	42,500	28	M
2	2015	50,000	29	M

Table 1 : Income of two people from 2013-2015

The information (the characteristics of income, age, and sex) gathered over the course of 3 years for various individuals is depicted in the table above. It displays the information gathered over a three-year period for two individuals (persons 1 and 2). (2013, 2014, and 2015).This is a typical Panel dataset table.

2. METHODOLOGY OF PROPOSED SYSTEM

The proposed system has models related to data collection, data pre-processing, defining the sensitivity of data, applying appropriate techniques to detect outliers and assessing the result produced by each of those algorithms in order to derive a conclusion. Understanding and modelling the available data are the core components of the outlier detection system.

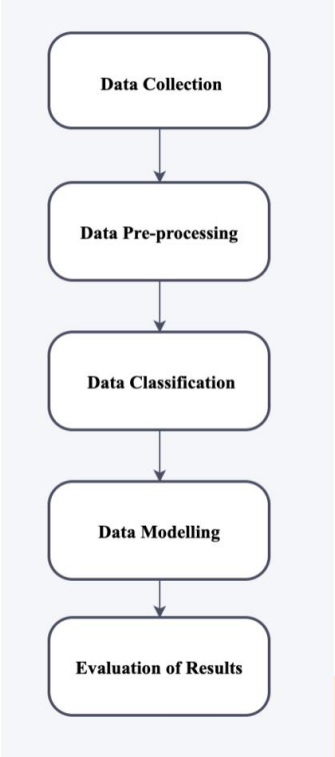

Figure 3:Methodology

2.1 Data Collection

Data collection is an important aspect in developing a proposed system. It is a systematic approach to accurately collect the information from various sources, the main driver to it is the quality of the data. This is a process of gathering, storing, visualising and analysing data from various sources. In this system we are more engrossed in time-series datasets. The data collection process has had to change and grow with time. Time Series data is a set of data points observed through repeated measurements over time, For example electrical activity in the brain over time, heartbeats per minute in a day, stock prices, annual retail sales and many more. Time series data can be of two types: measurements gathered in regular time intervals and irregular time intervals. Heartbeat measured per minute is measured in regular intervals and annual retail sales are measured in irregular intervals.

Time series data plays an important part in our daily lives as it treats time as the primary axis which is used for statistical analysis, forecasting and monitoring systems.

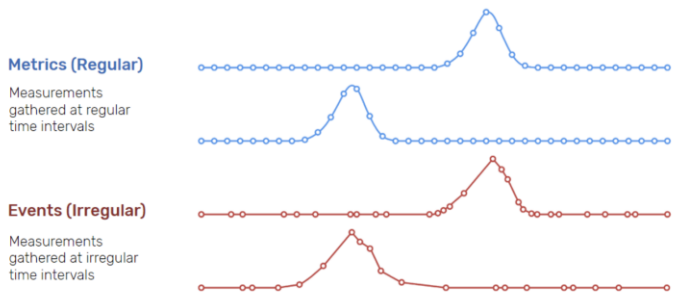

Figure 4: Classification of Time series data

Time series data are of two types univariate and multivariate. Univariate time series consists of a single observed value in regular or irregular time interval.Consider an example that shows the industrial index with respect to time -

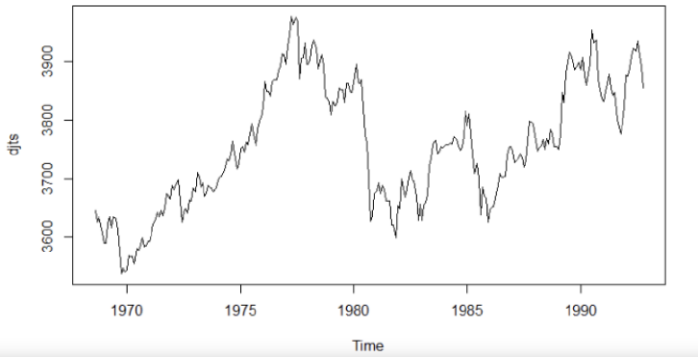

Figure 5: Univariate time-series data

This is univariate time series data that shows the industrial index value for different years from 1970 to 1990.

Multivariate time series data has data points in the datasets that have more than one time-dependent variable. If there are 2 variables that are dependent on time it's called Bivariate.

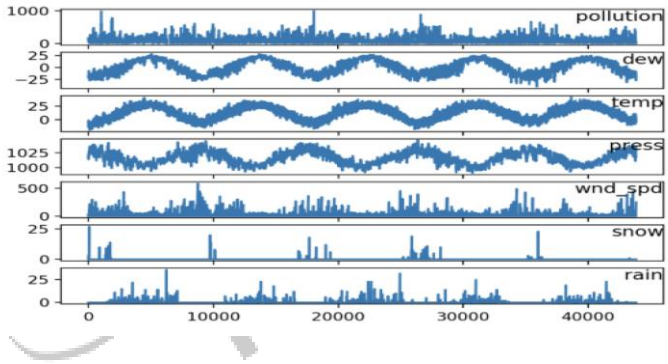

Figure 6: Multivariate time- series data

In this example we see that the air pollution is derived from all these variables and the time axis is based on the primary axis. Each variable has dependency on each other and all together can be used to predict air pollution.

2.2 Data Pre-processing

Data preprocessing refers to the procedures we must follow to alter or encode data so that a computer can quickly and readily decode it.The fundamental need for a model to make accurate and exact predictions is that the algorithm should be able to quickly understand the characteristics of the input.When data mining methods are used to this noisy data, the outputs would not be of high quality since the patterns would not be successfully identified. Therefore, data processing is crucial to raising the general level of data quality.

Here the proposed prepossessing model uses two method

2.2.1 Anova Test

By evaluating for variance-based variations in means between two or more groups, an ANOVA test is a type of demographic test used to evaluate if there is a statistically significant difference between them. The unconventional variable is divided into two or more groups by ANOVA. For instance, one group may be expected to have an effect on the dependent variable, whilst the other might be used as a control group and not be expected to have an effect.

The audacity for any parametric test apply to the ANOVA test are:

- The samples taken from the population should have a normal distribution.
- Case independence: The sample instances have to be separate from one another.
- The variance should be homogeneous, which implies that it should be about equal across all groups.

2.2.1 T-Test

T-Test is one of the statistical tests performed to compare the average or the means of two particular groups.The majority of the time, it is used to determine whether a procedure actually has an impact on the population of interest or to determine whether two groups differ from one another. The Anova test is used when many pairwise comparisons must be made because a T-Test can only be used to compare more than two groups. We may select from a variety of T-tests depending on whether we want to test the difference in a particular direction and if the groups being compared are from a single population or two separate populations. One-sample, two-sample, or paired tests are available.

- One sample test: When we want to compare a group against a standard value, we make use of one sample test.
- Two-sampled test: when we want to compare the groups from two different populations we make use of the two sampled test.
- Paired test: When we want to compare the groups from a single population, we use paired test.

2.3 Data Classification

Data is classified based on the sensitivity. It can be classified as highly sensitive, low sensitive and medium sensitive. Here the word sensitive basically means the confidentiality and the integrity of the data.

Highly sensitive data: They are those type of data which when destroyed in an unauthorised manner, will have a catastrophic effect on the organisation

Medium sensitive data: They are intended for only the purpose of internal use.

Low sensitivity data:These kind of data are intended for the use of public.These are low sensitive

2.4 Data Modelling

The proposed model aims in detecting the abnormal observations or the outliers in various kinds of data sets. The input given to the model is the dataset and the model works on what kind of outlier detection algorithm is best suitable for what kind of the dataset based on the sensitivity of the data which is simultaneously calculated by various statistical metrics such as the variance, standard deviation and covariance.

The inbuilt outlier detection algorithms such as the boxplot algorithm, Elliptic Envelope Algorithm, Isolation Forest Algorithm and Local Outlier factor algorithms are used as the basis for the proposed model.

Based on the accuracy of the algorithm for the particular dataset, it is said that particular algorithm is best suited for that dataset.

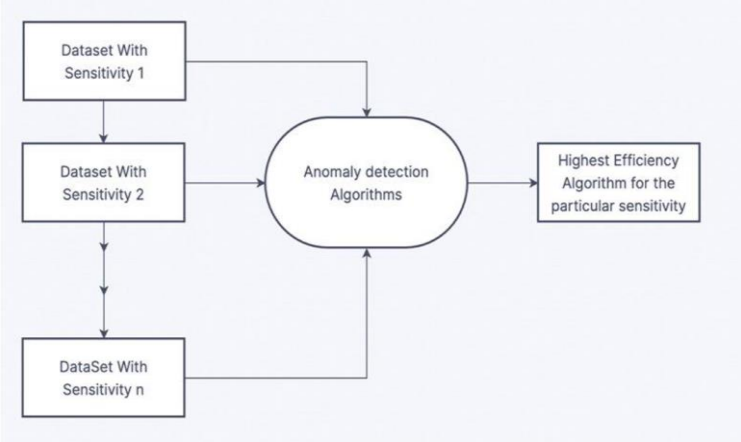

Figure 7:Data Flow Diagram

2.4.1 Box plot algorithm

Boxplot algorithm is the algorithm used for detecting the outliers in a univariate data.It enables a simpler way for the visualisation of the data.It also enables simpler comparison of characteristics of data within the categories.

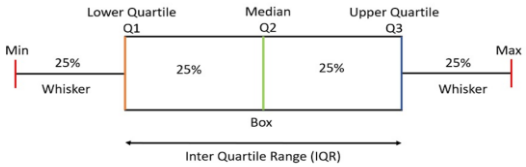

Figure 8:Illustration of Box Plot Algorithm

2.4.2 Isolation Forest algorithm

The isolation forest makes an effort to isolate each data point. In the case of 2D, it generates a line at random and tries to identify a point. Here, separating an anomalous point could just need a few steps, whereas separating closer typical points might require a lot more steps. Contamination is a crucial quantity in this context, and we should estimate its value based on trial and error while confirming our findings using outliers in a 2D plot. It represents the fraction of data points that are outliers. As this dataset only contains a few months' worth of data, Sklearn's Isolation Forest is

utilised; however, H2O's Isolation Forest, which is more scalable on large datasets, is also available and merits investigation.

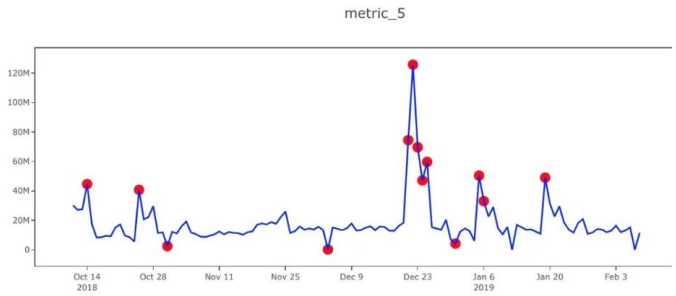

Figure 9:Illustration of Isolation Forest Algorithm in a graph

2.4.2 Local Outlier Factor (LOF)

An approach for unsupervised outlier identification is called the local outlier factor (LOF). Local outliers are points that are regarded as outliers based on their immediate surroundings. By taking into account the neighbourhood's density, LOF will detect an outlier. When the data density is not constant across the dataset, LOF works well.

By calculating the distances between nearby data points, local density is calculated (k-nearest neighbours). Thus, local density may be computed for each data point. By contrasting these, we may determine which data points have densities that are comparable to their neighbours and which have lower densities. Outliers are those having densities that are below average.

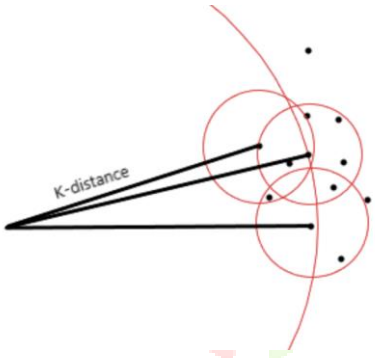

Figure 10:Demonstration of K-distance between points

2.5 Evaluation of Results

The accuracy of each algorithm with different types of sensitivity of the dataset are compared side by side.The result are to be illustrated as a table of performance of each of the dataset with one particular algorithm in order to find the method with is most suitable for that type if dataset sensitivity

3. Literature Survey

[1]Automatic Outlier Detection in Music Genre Datasets was given by Yen-Cheng Lu, Chih-Wei Wu, Chang-Tien Lu, Alexander Lerch. In this research paper they have done a comparative study of 6 outlier detection algorithms to a Music Genre Recognition. They have tried to test on how well the outlier detection algorithms can identify the mislabelled or corrupted music genre data record. On the basis of the results obtained they have tried to improvise the data set by applying various pre-processing techniques. They have extracted the features using the block-wise analysis method. They have also implemented the unsupervised clustering method K-means algorithm. They have used KNN-KNN method for defining the outlier scores of each data point by its distance to kth nearest neighbours.

[2]Outlier Detection in Traffic Data Set was by Teodora Mecheva. In this paper, they have compared the three outlier identification techniques namely the Local Outlier factor(LOF), Isolation Forest algorithm and Support vector machine over the traffic data sets.The data is obtained from the virtual detectors from the road cameras.A subset is made according to the hours of the day for a working day and they have applied all these algorithms over the each subset. The article also focuses on the purification of the dataset based on these three outlier techniques.They have calculated the efficiency of the methods by comparing the coefficient of variation of the raw data and the purified data. This task can be improvised by improving the methodology for data purification by extending the size of the dataset.

[3]Credit card fraud detection using outlier detection was done by S P Maniraj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed. In this article, the authors mainly aim at illustrating the modelling of dataset using machine learning with credit card dataset. This problem involves modelling the previously occurred transactions with the ones which have been detected to be fraudulent ones.With this modelling, the new transaction can be found whether to be a fraud transaction or not.The article has implemented Local Outlier factor and Isolation Forest algorithm for solving this problem.They have ensured that the project allows for various algorithms to work together. Here, the proposed system is implemented in Python. The Local Outlier factor provides the accuracy of 97% and the Isolation forest gives an accuracy of about 76%.

[4]Anomaly Detection in Clinical Data of Patients Undergoing Heart Surgery was done by Alva Presbitero, Rick Quax, Peter M. A. Sloot. The project aims in detecting and estimating the physical conditions of the patients to validate whether they are healthy or not.The anomalies in these kind of dataset might cause a fault in determining what disease the patient has. These anomalies are designated by the pattern changes. These pattern changes help in determining the transition from the healthy state to unhealthy state.They have made use of algorithms such as Isolation Forest Algorithm, Local Outlier factor algorithms to detect the anomalies in the dataset.The main drawback of the paper is that there is no specified technique to distinguish the critical patients from the non critical ones. As a result, important patients cannot be identified by simple outlier identification; instead, a broader strategy is needed, such as the surprise technique.

[5]Graph-based Anomaly Detection and Description was done by Leman Akola, Hanghang Tong, Danai Koutra. The project mainly aims on providing a structured overview on the state -of -art methods of outlier detection in data which can be represented by graphs.They have provided a general framework for the algorithms categorised under various categories such as Supervised vs. Unsupervised, Static vs. Dynamic approaches and for attributed vs. plain graphs. Graphs usually represent the interdependent nature of the attributes of the dataset.Different paths linking the data objects effectively capture the long range correlations.Besides the detection of the anomalies they also

concentrate on the anomalies which are detected and provide the survey of analysis of tools and techniques for post detection analysis and sense-making. The unsupervised clustering techniques such as Kmeans and agglomerative clustering are put into use in detecting the anomalies

[6]Outlier Detection : A Survey was done by Varun Chandola. The section has been divided into three main sections. In section 1 the problem statement is identified and in section 2,various aspects that constitute the exact formulation of the problem are identified. In section 3, the applications which deploy the outlier detection have been identified.Each data instance can be described using a set of attributes. The paper shows that the main challenge in any outlier detection algorithm is to identify the best set of attributes that allow the algorithm to calculate how efficient that particular algorithm, based on the parameter such as the accuracy.The importance of the outliers is generally represented numerically and they are categorised based on the level of importance. The categorization depicts the prioritisation of the outliers.

[7]Outlier Detection for Patient Monitoring and Alerting was done by Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper and Gilles Clermont.The paper interprets the effect of occurrence of anomalies in a medical dataset which could cause serious ill effects on the health of the patients.The total clinical coverage of alerting is enhanced by the outlier-based monitoring and alerting method. 4 486 post-cardiac surgery patients, data from the electronic health records, and a subset of 222 warnings produced by the method were used in the experiments. The outcome demonstrates a favourable correlation between statistical outliers and clinically significant warnings, supporting the idea that outlier-based alerting may have clinical applications.

[8]Anomaly Detection for Cyber-Security Based on Convolution Neural Network: A survey was done by Montdher Alabadi, Computer Engineering Department, Karabuk University, Karabuk, Turkey and Yuksel Celik Computer Engineering Department, Karabuk University, Karabuk, Turkey. The project aims on selection of approach for outlier detection for various kinds of domains.Distance based approach is used to determine the outliers by calculating the distances between the data objects with clear geometric interpretation.They have also calculate the “outlier factor” by defining a function F:x→R to characterise the outlier quantitatively.The function F basically depicts the amount of distance between the given object x and the other objects R in the dataset.The density based approach is also used to determine the abnormalities that lies in the neighbourhood of the data objects.Thus, the identification of anomalies actually helps to identify the traffic attacks based on its deviations from the already established profiles of data abnormality.

[9]Outlier Detection in Network Traffic Monitoring by Marcin Michalak , Łukasz Wawrowski , Marek , Rafał Kurianowicz, Artur Kozłowski and Andrzej Białas. This paper present results presented on the real outer traffic data which was collected in institute.The paper mainly focuses on two variables namely the number of data packets and the size of the packets.Hence the dataset is bivariate and the outlier is required to be checked in two dimensional space.The experiments were carried out in two modes. The first approach aims at learning the gathered data in the best way possible.The second approach aims at how these methods can be practically applied in real time applications.

4.Expected Outcome

One of the most crucial preprocessing procedures in data analytics is outlier detection, which is also thought to be essential for machine learning algorithms' optimal performance. In this paper, many approaches are discussed while keeping in mind the requirement for a reliable and simple outlier identification system.

The newly presented solutions are based on innovative statistical methods that are applied to different types of sensitivity of datasets.The outcome here concludes which algorithm would provide the highest efficiency and accuracy for the specified sensitivity of the provided dataset.

REFERENCES

[1] Yen-Cheng Lu, Chih-Wei Wu, Chang-Tien Lu, Alexander Lerch. Automatic Outlier Detection in Music Genre Datasets. KNN-KNN implementation. August 2016 DOI:10.1085/9.1093854 PMID: 356087659.
[2] Teodora mecheva outlier detection in traffic data set. 17th international conference on concentrator photovoltaic systems (cpv-17) DOI:10.1063/5.0093554 [3]S P Maniraj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed. Credit card fraud detection using outlier detection October 2021. International Journal of Scientific and Technology DOI:10.17577/IJERTV8IS090031
[4] Alva Presbitero, Rick Quax, Peter M. A. Sloot . Anomaly Detection in Clinical Data of Patients Undergoing Heart Surgery. December 2017 DOI:10.1016/j.procs.2017.05.002
[5] Leman Akoglu , Hanghang Tong, Danai Koutra Graph-based Anomaly Detection and Description: A Survey. June 2022 DOI: 10.18280/ts.390327
[6]Outlier Detection A Survey by Varun Chandola, University of Minnesota Arindam Banerjee, University of Minnesota and Vipin Kumar, University of Minnesota.
[7] Outlier Detection for Patient Monitoring and Alerting by Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper and Gilles Clermont.
[8] Anomaly Detection for Cyber-Security Based on Convolution Neural Network: A survey by Montdher Alabadi Computer Engineering Department, Karabuk University, Karabuk, Turkey and Yuksel Celik Computer Engineering Department, Karabuk University, Karabuk, Turkey
[9] Outlier Detection in Network Traffic Monitoring by Marcin Michalak , Łukasz Wawrowski , Marek , Rafał Kurianowicz, Artur Kozłowski and Andrzej Białas
Research Network Łukasiewicz, Institute of Innovative Technologies EMAG, ul. Leopolda 31, 40–189 Katowice, Poland