2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) | 978-1-7281-9675-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDABI51230.2020.9325683

Incremental Outlier Detection in Air Quality Data Using Statistical Methods

Manish Mahajan
Computer Science Engineering Department
Graphic Era Deemed to be University
Dehradun, India
https://orcid.org/0000-0002-5670-650X

Santosh Kumar
Computer Science Engineering Department
Graphic Era Deemed to be University
Dehradun, India
https://orcid.org/0000-0002-1008-0804

Bhasker Pant
Computer Science Engineering Department
Graphic Era Deemed to be University
Dehradun, India
https://orcid.org/0000-0002-6070-5104

Umesh Kumar Tiwari
Computer Science and Engineering Department
Graphic Era Deemed to be University
Dehradun, India
https://orcid.org/0000-0003-2837-9334

Abstract—**Air pollution is one of the biggest problems being faced by the world today and effective measures are being sought to counter this menace. In order to devise a proper strategy of countermeasures, it is imperative to ensure that the data on which the modeling and decision making is being based is clean. Outliers usually creep in the sensor generated data due to human and system errors and have a degrading effect on the decision making. A number of methods have been tested for detection of outliers but an incremental approach is highly sought after as the temporal air quality data is available as a data stream. We have proposed a framework to analyze the performance of statistical outlier detection methods as being the answer to the continuous analysis problem. We have presented a comparative analysis of five statistical methods viz. Z-score, InterQuartile Range, Grubb's Test, Hampel's test, Tietjen-Moore Testfor outlier detection when used over the complete dataset and when used in an incremental mode.**

Keywords—Outlier, Air quality data, Statistical methods, Incremental mode

I. Introduction

The day to day activities that we engage in causes air pollution, which is currently the most pressing problem being faced by the world. Large cities which are prone to ever rising levels of industrialization and growing urbanization are especially susceptible to the menace of air pollution [1]. The rising levels of air pollution results in serious health hazards leading to decreased quality of life. A number of ailments currently plaguing the globe can be attributed to the rising levels of air pollution [2]. The methods needed to counter the effects of air pollution need accurate and correct data about the pollutants present in the air. Air Quality Index (AQI) is calculated and used by the environmental monitoring agencies to communicate the quality of air and the level of pollution in the air. For example, high AQI values are used to recommend reduced outdoor activities as it indicates poor air quality [3]. Typically air pollution is measured by measuring the concentration of the harmful pollutants in the air and using these readings to calculate the AQI. Air quality research relies very heavily on the data provided by the surface sensors as these surface pollutant measurements provide crucial information which is imperative for research and decision making by helping to formulate and test the models [4]. The major pollutants that are considered the main cause of air pollution and are most commonly taken into account for air quality measurement include:

- Sulpher Dioxide (SO_2)
- Nitrogen Oxides (NO_x)
- Ozone (O_3)
- Particulate Matter (PM_{10} or $PM_{2.5}$)
- Carbon Monoxide (CO)
- Carbon Dioxide (CO_2)

The quality of data is vital for efficient interpretation and decision making and this is also true for the air quality research. In spite of all the quality assurance and quality control measures in place, outliers can still creep in. the main reasons for this could be instrument malfunction, incorrect placement, limitations faced by measurement methods or environmental conditions [5][6].

Outliers can either be boons or banes depending on how the application area looks at these deviations. Applications like medical diagnostics, fraud detection and intrusion detection specifically look for the outliers as these indicate deviant behavior that is of interest to such application areas. On the other hand, application areas like air quality analysis, demographic analysis etc. have the quality of decision making hampered by the outliers and are benefitted by elimination of outliers [7]. Outliers have a severe detrimental effect on the quality of models being developed using such data and subsequently on the decision making by employing these models But in both the cases, the main problem to be addressed is the detection of outliers.

Over the years, many methods have been tried for the effective detection of outliers, with most of the research being focused on the data mining based methods including Distance based, Density based methods etc. [8]. Such methods, although providing a good level of accuracy are limited by their requirement of complete data set and also being resource intensive. The requirement of resources often proves to be an hindrance in setting up of such countermeasures in developing and third world countries, where these are most needed. The temporal data streams that are received from sensors measuring air quality data need to be processed incrementally and in real time to be effective in decision making [9]. Statistical methods have the advantage in this aspect as they are fast, almost equally accurate and can be processed even with limited resources. Statistical methods of outlier detection usually work on the premise that they calculate some statistical measure from the data available and designate the data item as being normal or outlier based on this measurement [10]. This measurement is incrementally updated as more of the data becomes available. Thus statistical methods are capable of handling the air quality or any other type of temporal data in an incremental manner thereby eliminating the requirement of needing complete data set before modelling and decision making. This capacity is very useful in applications using sensor generated data as the data comes as a time series and can be continuously updated and utilized [11].

This paper evaluates the comparative performance of several statistical methods of outlier analysis being used incrementally. Their performance is compared to the performance of the same methods when complete data series is available beforehand. The paper is organized as: Section 2 discusses the different statistical methods of outlier detection along with the current research in the areas; Section 3 discusses the proposed research framework and methodology; Section 4 elaborates the findings from applying these methods.

II. Methods Tested in the Current Research

Graphical methods like Boxplots, Histograms and scatterplots are the easiest tools for visually detecting the most obvious outliers. Histograms can help in visually identifying extreme value outliers that are way different from the normal trend of values. But this method fails in identifying the subtle outliers as well as the contextual outliers. Boxplots use one of the most basic methods of flagging the outliers, that is the *Inter Quartile Range* method [12].

The scatterplot along with the regression line is another simple way of identifying both univariate and multivariate outliers visually but is limited by its inability to efficiently identify contextual outliers and also fails to differentiate between outliers and novelties thereby resulting in a much higher False Positive rate. Most of the established methods used for identifying outliers to be eliminated for the purpose of improving the accuracy and efficacy of the model are based on calculating some statistical measure and then updating it incrementally as more of the temporal data becomes available. The data item is compared against this measure and is flagged as a normal or an outlier value accordingly [13].

A. Z-score

Z-scores are an easy and fast method of identifying anomalous values but have the limitation that they assume a Gaussian distribution of data. It is a parametric outlier detection method. Z-scores essentially measures how many numbers of standard deviations above and below the mean does the data item fall. The outliers are the data points that lie in the two tails of the distribution and are therefore at a greater distance from the mean [14][15]. In its basic form the Z-score is calculated as

$$Z_i = \frac{x_i - \mu}{\sigma}$$

Where x_i is a data point, μ is the mean of all x_i and σ is the standard deviation of all x_i. A threshold is defined for the Z-score based on the application and type of data to flag the outliers.

$$|Z_i| > Z_{thr}$$

A Z-Score of more than +/-3 usually flags an errant value. The Z-score in its original form is not very efficient in identifying the outliers as the value of Z itself is influenced by the presence of outliers as they can inflate the central tendency measures used [16][10].

Modified Z-Score method [17] replaces the mean and standard deviation used in the Z-score with the median and the Median Absolute Deviation (MAD) to make it more resilient towards effect of outliers. Modified Z-score is defined as

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Where $\tilde{x}$ is the median of the data.

B. InterQuartile Range

The InterQuartile range(IQR) can be used in combination with an adjustment factor to build boundary fences that can help to identify the extreme value outliers. The outlier is the data point x_i that lies outside the IQR [9].

$$x_i > Q_3 + k(IQR) \ OR \ x_i < Q_1 - k(IQR)$$

$$where \ IQR = Q_3 - Q_1 \ and \ k \geq 0$$

The IQR, being a measure of statistical dispersion, is much more resilient towards the presence of outlier values in the data set. The IQR is the difference between the upper and lower quartiles. The value k is chosen based on the application and characteristic of data to filter out the outliers. Typically k is chosen as 1.5 [15][18].

Boxplots are an extension of the IQR method and are used to represent the findings graphically and to visually identify the outliers. A typical Boxplot showing the placement of fences and outliers are shown in figure 1.

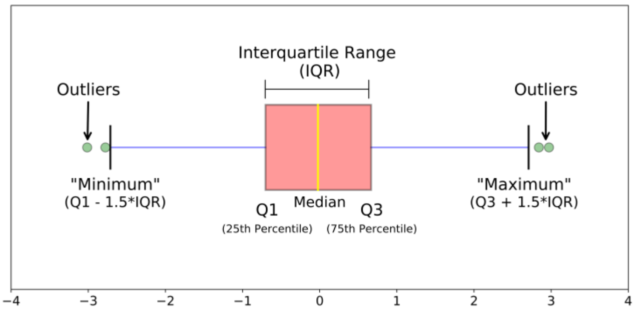

Fig. 1. Sample Boxplot showing Outliers

C. Grubb's Test

Grub's Test requires the data to be based on normal distribution just like Z-score method. This method can be applied to the data set incrementally till all the outlier values are identified [19].

Grubb's test is essentially a test of hypothesis and defines the two hypotheses as –

H_0 *assumes the absence of any outliers in the data set*

H_1 *assumes presence of at least one outlier in the data set*

The Grubb's test is implemented as

$$G = \frac{\max_{i=1\ldots N} |x_i - \mu|}{\sigma}$$

Wherex_i is a data point,N is the number of data items, μ is the mean of allx_i and σ is the standard deviation of all x_i [14][20].

The calculated G value is compared against the critical value to distinguish the outlier data items. The test also requires establishment of a proper level of significance, which indicates the maximum error level (False Positive and/or False Negative) that is acceptable, and is highly dependent on the application type and data type [13].

D. Tietjen-Moore Test

The Tietjen-Moore test is a generalization of the Grubb's test and is used to detect the presence of multiple outliers in a univariate data set that is distributed approximately by a normal distribution. The test requires the number of outliers to be specified. If the test is used to check a single data point for outlier, it is similar to the Grubb's test [21]. The test requires the data set to be sorted and then the test statistic for the k largest points (where k is the approximate number of outliers) is calculated as

$$L_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where$\bar{y}$ is the sample mean for the full sample under test and $\bar{y}_k$ is the sample mean of the dataset from which the k largest values have been removed. The test statistic for the k smallest points (where k is the approximate number of outliers) is calculated as

$$L_k = \frac{\sum_{i=k+1}^{n}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where$\bar{y}$ is the sample mean for the full sample under test and $\bar{y}_k$ is the sample mean of the dataset from which the k smallest values have been removed [22][18].

Simulation is an important aspect of the test as it is used to determine the critical region. The test is always a lower one-tailed test [23][20].

E. Hampel's Test

Hampel's test is quite resilient to the presence of outliers in the dataset and is least affected by them. It is also suitable for small as well as large data sets. The test uses the median value from the test dataset $\tilde{x}$ as the value separating the two halves of the dataset. Next the test calculates the deviation of each value of the dataset from the median value [20].

$$r_i = (x_i - \tilde{x})$$

The median from the deviation is identified $\widetilde{r_i}$ and is used to check for outliers [14].

This is done by checking the condition

$$|r_i| \geq 4.5\widetilde{r_i}$$

A data point that exceeds the condition is flagged as an outlier [10][13][15].

III. Proposed Framework and Methodology

The KNIME Analytics platform was used to actually implement the various outlier detection methods and to perform the tests [24]. The framework was initially tested on the sample air quality datasets obtained from UCI Library and Outlier datasets obtained from the ELKI dataset collection. Later the framework was applied on the actual air quality data for the different cities in India. The datasets used for the actual experiments covered the air quality measurements as obtained from various sensor stations across the country over a period of 2010 to 2015. The effort of outlier detection was focused on the particulate matter concentration, both PM10 and PM2.5 as they are considered to be the most hazardous to human health and are thought to be causing the maximum damage. To make the comparison between the incremental and non-incremental mode of operation of the various tests, a sliding window technique was used. Different window sizes were used to ensure that the results obtained were not coincidences. Figure 2 shows the proposed methodology used for carrying out the comparisons.

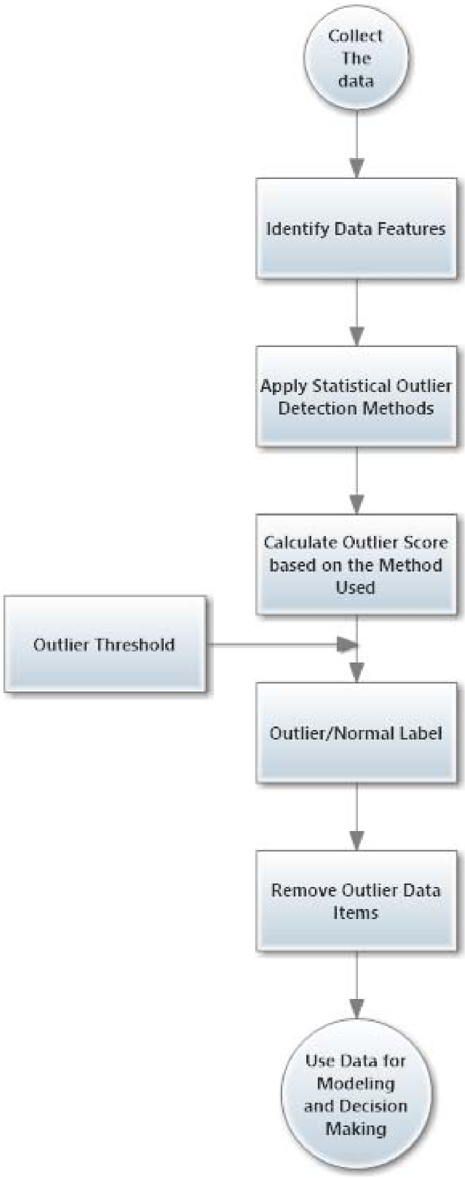
Collect The data
Identify Data Features
Apply Statistical Outlier Detection Methods
Calculate Outlier Score based on the Method Used
Outlier Threshold
Outlier/Normal Label
Remove Outlier Data Items
Use Data for Modeling and Decision Making

Fig. 2. Proposed framework

IV. Experimental Observations and Results

While using the various methods with the sample Outlier datasets from the UCI and ELKI datasets, the various methods were evaluated in terms of the fraction of outliers identified and the time taken by the tests. Figure 3 shows the results for 5 of the various sample datasets in the terms of percentage of outliers identified where the number of the outliers was known beforehand.

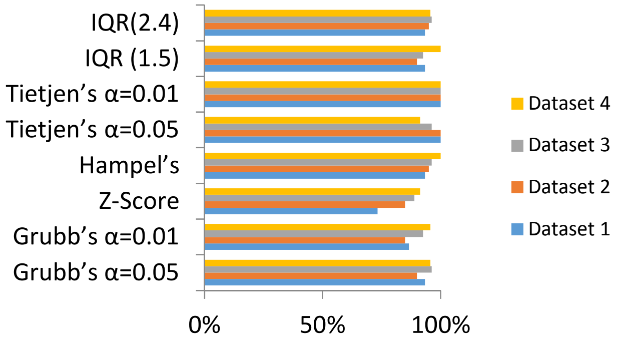

Fig. 3. Performance of Methods

The various methods have also been compared based on the time taken by them to identify these outliers. Figure 4 shows the results for the same sample datasets in terms of time taken by them to identify the outliers. To ensure correctness of the time measurement, five measurements were made.

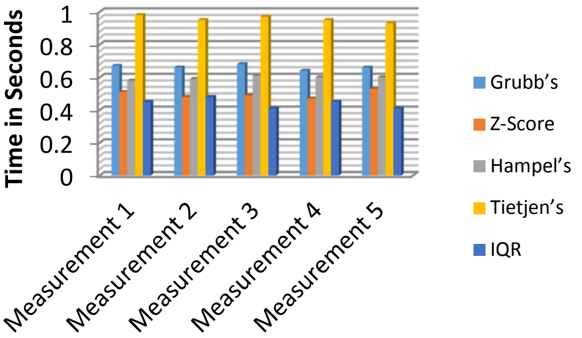

Fig. 4. Time taken by the various methods

As is clear from the experiments conducted on the sample datasets, the IQR and the Hampel's tests were able to perform quite well in terms of giving adequate outlier detection performance and having the lower time requirements. The outlier finding accuracy was slightly better in Tietjen's test but the difference was not significant and was offset by the fact that the it required the most time.

Another comparison has been made between the performance of the various tests when applied in an incremental mode and when applied with the entire set of data. Sliding window sizes of 5%, 10%, 15% and 20% of the entire sample dataset sizes were used for testing. The results are shown in figure 5.

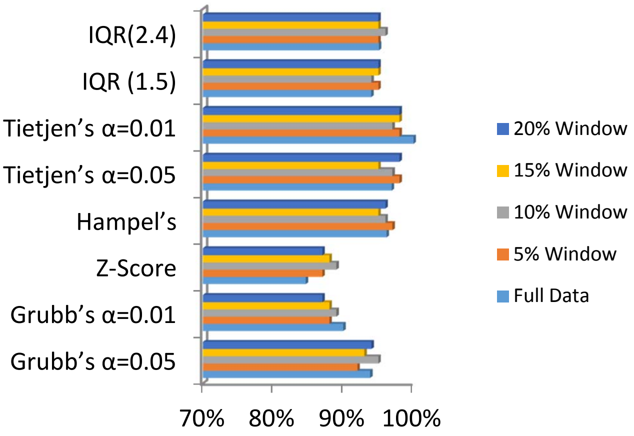

Fig. 5. Comparison of incremental and full approach

The results shown in the figure 5 also supported our hypothesis that the statistical methods don't show any significant degradation when used in an incremental manner and in fact in some cases, using them in an incremental manner actually improved the performance.

Encouraged by the results, the methods were then applied on the actual air quality data sets available for the different states of India. A total of 10 states were chosen for testing based on the number of samples available. Although the outliers were not pre-identified here as was the case of the sample datasets but the results of the application of the methods to sample datasets helped build the confidence that the methods would still perform well in this scenario also. Table 1 shows the outlier identification statistics of the various methods used in both incremental and non-incremental mode by using the IQR method. The table lists the number of outliers identified from the datasets of various states by using the IQR method. The results have been shown for the test applied on the entire data as well as the test applied in incremental sliding window fashion using four different window sizes as percentage of the complete data set.

TABLE I. No. of Outliers identified by the IQR Test

State	Size of data	Full Data	Window			
			5%	*10%*	*15%*	*20%*
Maharashtra	60384	58	57	56	57	58
Uttar Pradesh	42816	47	44	48	47	46
Andhra Pradesh	26368	35	35	34	34	34
Punjab	25634	34	32	34	33	33
Rajasthan	25589	38	37	36	36	36
Kerala	24728	22	21	22	20	22
Himachal Pradesh	22896	62	60	60	62	61
West Bengal	22463	45	44	43	43	45
Gujarat	21279	32	32	32	30	30
Tamil Nadu	20597	27	25	26	27	25

Table 2 shows the number of outliers detected from the same states dataset but this time using the Hampel's Tests.

TABLE II. NO. OF OUTLIERS IDENTIFIED BY THE HAMPEL'S TEST.

State	**Size of data**	**Full Data**	**Window**			
			5%	***10%***	***15%***	***20%***
Maharashtra	60384	59	57	59	58	60
Uttar Pradesh	42816	46	47	47	44	46
Andhra Pradesh	26368	33	32	31	31	33
Punjab	25634	35	36	34	33	35
Rajasthan	25589	38	36	38	38	36
Kerala	24728	22	23	22	21	21
Himachal Pradesh	22896	60	58	61	59	59
West Bengal	22463	44	42	44	43	45
Gujarat	21279	32	32	32	31	31
Tamil Nadu	20597	27	27	25	27	25

The results clearly show that there is little or no loss of precision in the performance of the methods. In fact in some datasets and some statistical methods, it actually leads to an improvement of results.

V. CONCLUSION AND FUTURE WORK

The results from both the sample and real life datasets have established the fact that the statistical tests are capable of being applied in an incremental manner and therefore have the capability of analyzing the stream of data as coming from the air pollution sensors. It is clear from the difference in performance while using the methods in an incremental fashion and using them on the full data is not more than 4%. This is more than compensated by the advantages gained out of using them in the incremental manner. Using the methods in an incremental mode has the added advantage that the outlier detection can be used in "real time" and can affect the decision making as more and more data is received from the sensors. The statistical methods are not much resource intensive and hence can also be implemented in low-cost analysis centers, thereby reducing the infrastructure investment which is very important for Developing countries which are in most need of such air pollution countermeasures. The air quality data is generally low dimensional in nature, having at most 6-10 attributes.

The statistical methods need to be evaluated for higher dimension data to ensure their efficiency in handling such data. Also the incremental calculation of the statistical measure needs to be checked for high volume and high speed data.

REFERENCES

[1] R. Burnett *et al.*, "Global estimates of mortality associated with longterm exposure to outdoor fine particulate matter," *Proc. Natl. Acad. Sci. U. S. A.*, 2018.

[2] A. J. Cohen *et al.*, "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015," *Lancet*, 2017.

[3] K. Balakrishnan *et al.*, "The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017," *Lancet Planet. Heal.*, 2019.

[4] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain, "On the feasibility of measuring urban air pollution by wireless distributed sensor networks," *Sci. Total Environ.*, 2015.

[5] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*. 2017.

[6] S. De Vito, M. Piga, L. Martinotto, and G. Di Francia, "CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization," *Sensors Actuators, B Chem.*, 2009.

[7] C. C. Aggarwal, *Outlier Analysis*. Cham: Springer International Publishing, 2017.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*. 2009.

[9] H. Wu *et al.*, "Probabilistic Automatic Outlier Detection for Surface Air Quality Measurements from the China National Environmental Monitoring Network," *Adv. Atmos. Sci.*, 2018.

[10] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2018.

[11] H. Yao, X. Fu, Y. Yang, and O. Postolache, "An incremental local outlier detection method in the data stream," *Appl. Sci.*, 2018.

[12] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 2005.

[13] C. C. Aggarwal and C. C. Aggarwal, "Probabilistic and Statistical Models for Outlier Detection," in *Outlier Analysis*, 2017.

[14] M. Abzalov, "Exploratory data analysis," in *Modern Approaches in Solid Earth Sciences*, 2016.

[15] O. Schabenberger and C. A. Gotway, *Statistical methods for spatial data analysis*. 2017.

[16] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate Data Analysis," *Vectors*. 2010.

[17] H. Aguinis, R. K. Gottfredson, and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers," *Organizational Research Methods*. 2013.

[18] D. Machiwal, M. K. Jha, J. Harris, J. C. Loftis, and R. H. Montgomery, "Comparative evaluation of statistical tests for time series analysis: application to hydrological time series Statistical methods for characterizing ground-water quality ," *Hydrol. Sci. Journal/Journal des Sci. Hydrol.* , 2008.

[19] F. Angiulli, F. Fassetti, G. Manco, and L. Palopoli, "Outlying property detection with numerical attributes," *Data Min. Knowl. Discov.*, 2017.

[20] S. S. Tripathy, "Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution," Am. J. Theor. Appl. Stat., 2013.

[21] N. B. Aissa and M. Guerroumi, "Semi-supervised Statistical Approach for Network Anomaly Detection," in Procedia Computer Science, 2016.

[22] D. L. S. Reddy, M. Ramchander, B. R. Babu, and M. Geetalatha, "Comparitive study of outlier analysis methods in improving classifier accuracy on categorical data," in International Conference on Microelectronics, Computing and Communication, MicroCom 2016, 2016..

[23] C. E. Efstathiou, "Estimation of type I error probability from experimental Dixon's 'Q' parameter on testing for outliers within small size data sets," *Talanta*, 2006.

[24] KNIME.COM AG, "KNIME Analytics Platform," *KNIME Analytics Platform product sheet*, 2016.