ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 16-21
doi: 10.18178/wcse.2020.02.004

Analysis of Outlier Detection on Structured Data

Khin Myo Myat[1+] and Si Si Mar Win[1]

[1]University of Computer Studies –Mandalay, Myanmar

Abstract. Outlier detection has played an important role in all research areas for data analysis in various domains. Outlier is involved according to human error, sensors or mechanical faults and environment. It is detected to get data quality for all applications. On the other hand the good outliers can occur by chance in new contributions in research .The aim of this study is to discover new trend with outlier in dataset. The problem statement is how to detect the outlier analysis based on different datasets between before and after removing outlier. In this system, a box and whisker plot and the robust J48 algorithm are applied for outlier detection and classification.

Keywords: a box and whisker plot, classification, data Mining, J48, outlier detection

1. Introduction

Today the technology revolution has enabled us to gather massive amounts of data from different sources like social networks, sensor data, scientific data, biological data and networked systems data, etc. In real world data are incomplete, lacking attribute values and containing errors or outliers. So it will be important to have access to reliable data to make the decision and data preposition plays an important role in machine learning. Scientific experiments are especially sensitive situations when dealing with outliers. Outliers mean data points that are far from other data points. A data point is a discrete unit of information. Any single fact is a data point. The outlier can be a result of a mistake during data collection or it can be just an indication of deviation in the data. Outlier detection is an interesting problem in machine learning with the application from weather, computer security, financial and medical research domain.

In data mining, anomaly detection is the identification of unusual items, event or observations which raise uncertainty by differing obviously from the majority of the data. Unsupervised anomaly detection techniques detect anomalies in an unlabelled test dataset under the assumption that the majority of the instances in the dataset are normal by looking for instances that seem to fit least to the remainder of the dataset. In this paper, the different datasets and the WEKA toolkit were used to compares the accuracy which dealing with outliers and omitting outliers of data set. The box and whisker plot is used to outlier analysis on the different data set. It is needed around decisions why a specific data instance is or is not an outlier.

The paper is organized accordingly the introduction of outlier detection of data in section1. And then section2 describes the related works based on research paper. In section 3 methodologies is also described in section 4, the proposed system has been stated .In section 5, experimental results & analysis. And finally concludes with future work.

2. Related Work

In this section, the related work is presented in terms of outlier detection in preprocessing and classification of data mining. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will

[+] Corresponding author. Tel.: +95-9971323888
E-mail address: khinmyomyat558@gmail.com

be considered abnormal. Before abnormal observations can be signed out, it is necessary to characterize normal observations.

Classification has been successfully applied to a wide range of application areas, such as scientific experiments medical diagnosis, weather prediction, credit approval customer segmentation, target marketing and fraud detection [1, 2]. Decision tree classifiers are used extensively for diagnosis of breast tumor in ultrasonic images, ovarian cancer

Heart sound diagnosis and so on Arvind Sharma and P.C. Gupta discussed that data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [3].As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed. Taking into account the prevalence of diabetes among men and women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool [4].

Data in most large databases is not perfect. The data called outliers or noise highlights a pattern that does not coincide with the pattern shown by the majority of data. So these data should be excluded from regular processing of data mining. Therefore, outlier detection (also known as anomaly detection) is required to find the outliers to improve the quality of data and to acquire a more accurate result of data mining. Outlier detection is an observation that deviates so much from other observations so as to arouse suspicions that it was generated by different mechanism. Outliers are probably generated because of measurement or executed error , etc., at the same time it could be considered as noise generally or decreased its effect in view of revised outlier value in order to pretreated data set.

R.Delshi Howsalya Devi et al.suggested a novel hybrid outlier detection based data mining algorithm to find the outliers in test data [5].

Data in most large databases is not perfect due to noise or outliers. . Removing outliers improves the quality of data and to acquire a more accurate result. Zheng et al. applied outlier detection in preprocessing step on the large cancer dataset. [6].

Some authors described the approach to detect anomaly without knowing the knowledge of anomaly class using an anomaly validation set for selecting hyper parameters of one class RBF-SVMs.They proposed for anomaly detection using classifiers to perform features to be a late fusion of hidden layer activation and residual error vectors and raw input signals.[7].

Recently, organizations have concluded that processing big data, especially the data coming from Twitter and Facebook can provide a significant impact on increasing the business's effectiveness and added values [8, 9].

Thomas et al. presented a numerical scheme for generating anomaly detection model that reduces to fitting distributions to data. They detected the problem of providing accurate ranking of disjoint time periods in raw IT system by monitoring data with their anomalousness. They suggested that their methods can be used to analyze various types of anomaly datasets. [10].

The data with outliers may also reduce clustering quality. So the author proposed a new outlier detection method to select initial cluster centers based on data density for adaptive K- Means Method [11].

The special characteristics of big data streams, such as transiency, uncertainty, multidimensionality, dynamic relationship, and dynamic data distribution, introduce new problems that make outlier detection for big data Streams more challenging [12]

3. Methodology

Data mining is a relatively a new technique to the world of information science. Detecting outliers, instance in database with unusual properties, is the data mining task. Now Outlier detection is the most researchable area in data mining field for knowledge discovery. An outlier is a rare chance of occurrence

within a given data set. There are good outliers that provide useful information that can lead to the discovery of new knowledge and bad outliers that include noisy data points.

The current section describes four main categories which include outlier detection in pre-processing, box and whisker plot, j48 algorithm for classification, Data Source and tools for experiments.

3.1. Outlier Detection

Outlier detection and analysis is an important data mining problem that goals to get anomaly points and behavior in data sets. It may be defined as the process of detecting and subsequently excluding outliers from a given dataset. To do so, outlier detection is vital to get the high quality of dataset. There are no standardized Outlier a branch of data mining has many applications in data analysis. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly.

Unlike many other methods of data display, they show outliers within a dataset. Outlier detection and analysis is an important data mining problem that goals to get anomaly points and behaviour in data sets. It may be defined as the process of detecting and subsequently excluding outliers from a given dataset.

3.2. Box and Wisher Plot

Box plots are useful for comparing datasets, especially when the datasets are large or when two or more data sets are being compared and when they have different numbers of data elements. It is a standardized way of displaying the distribution of data based on a five number summary. They are the minimum, first quartile Q1, median, third quartile Q3 and maximum. There are a few important vocabulary terms in a box and whisker plot method.

- Q1-quartile 1, the median of the lower half of the data set.
- Q2-quartile 2, the median of the entire data set.
- Q3-quartile 3, the median of the upper half of the data set.

IQR-Interquartile range, the difference from Q3 to Q1, which is the width of the box in the box and whisker plot. Extreme values- the smallest and largest values in a data set. In order to be an outlier, the data value must be larger than Q3 by at least1.5 times the interquartile range IQR or smaller than Q1 by at least 1.5 times the IQR when two or more data sets are being compared.

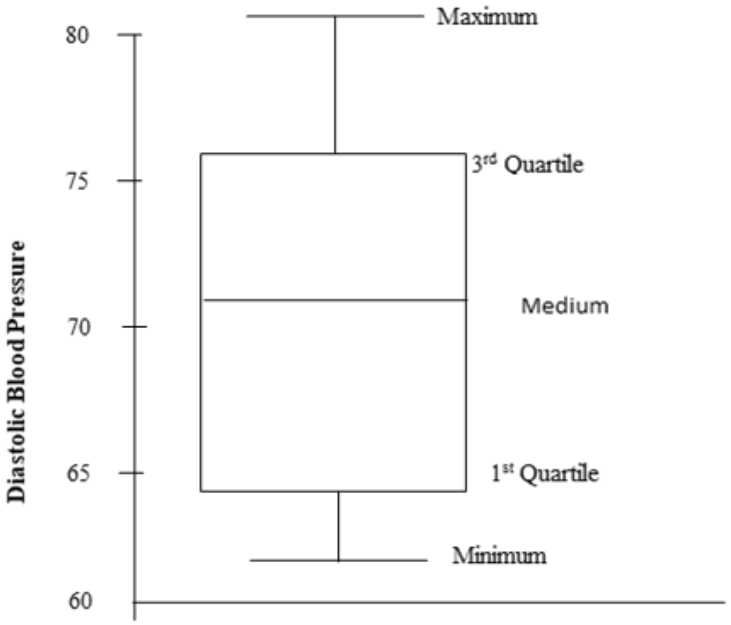

Fig. 1: Box and whisker plot

3.3. J48 Algorithm

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48

is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for pr écising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

3.4. Data Source and Tool

We tested the variety of datasets on data mining tool such as WEKA. In this paper the WEKA toolkit were used to calculate the accuracy with outliers without outliers. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can be used directly to the diabetes dataset. WEKA consists of tools for data preprocessing, classification, regression, clustering, association rules and visualization. It is open source software issued under the General public License. It is implemented in the java programming language and runs in any modern computing platform. It is portable and platform independent .it provides a graphical user interface for exploring and experimenting with machine learning algorithms on the datasets.

4. Proposed System

The proposed system with anomaly detection system detects outlier in the datasets by outlier detection method based on interquartile range and classification is done by using J48 classifier. In this paper, the comparison of accuracy which dealing with outliers and omitting outliers of dataset is presented. This paper discusses about the concept of outlier and outlier detection approaches. In this paper a combined approach such as a box and whisker for outlier analysis and J48 classifier for evaluating the effectiveness of outlier detection process. At first, a box and whisker plot algorithm is used for detecting outliers in the dataset. After the outliers have been removed, the data are given as input into a J48 classifier to evaluate the accuracy of dataset. Among all the classifiers, J48 is one of the best classifier in data mining.

By making a box and whisker plot Learn how to recognize potential outliers. Here, Diastolic blood pressure (mmHG) attribute in Diabetes dataset is considered for outlier detection. The values are 72.0, 66.0, 64.0, 66.0, 40.0, 74.0, 50.0, 500.0, 70.0, 96.0, 92.0, and 74.0

1. Arrange all data points from the lowest to highest
   40.0, 50.0, 64.0, 66.0, 66.0, 70.0, 72.0, 74.0, 74.0, 92.0, 96.0, 500.0
2. Calculate the median of the dataset.
   The median Q2= $\frac{70+72}{2}$ =71
3. Calculate the lower quartile Q1 is the data point below which 25% of the observations set.
   The lower quartile Q1=$\frac{64+66}{2}$ =65
4. Calculate the upper quartile Q3 is the data point above which 25% of the dataset.
   The upper quartile Q3=$\frac{74+92}{2}$ =83
5. Find the interquartile range
   Q3-Q1=83-65=18
   Lower limit = Q1-(1.5*IQR) =38
   Upper limit = Q3+ (1.5*IQR) =110

An outlier in example, it would have to be less than 38, which is the difference between Q1 (65) and IQR (18).Similarly, it would have to be more than 110, which is the adding Q3 (83) and IQR (18).

The values which are beyond these are extreme greater than 110 is 500 outlier.

The mean of our dataset with outliers is
(72.0+66.0+64.0+66.0+40.0+74.0+50.0+500.0+70.0+96.0+92.0+74.0)/12=179.66

The mean of our dataset without outliers is
(72.0+66.0+64.0+66.0+40.0+74.0+50.0+70.0+96.0+92.0+74.0)/11=69.45

Since the outlier can be attributed to human error and because it's inaccurate to say that the blood pressure was almost 180 mmHG, so should omit to this outlier

5. Experimental Result

Experiment are carried out on WEKA with 10 fold cross validation is where a given dataset is split into a 10 number of folds where each fold is used as a testing set at some point. It is used to perform statically analysis of the individual attributes in dataset. In this paper, the performance indicators such that RMSE, ROC, accuracy, Precision, Recall based on true positive and false positive are compared for the dataset using J48 algorithms and outlier detection algorithms. By comparing the accuracy and correctly classified attributes, suitable decision can be figure out.

The data sets within WEKA experiment are used to evaluate an attribute's efficiency by considering its mean, min, max, standard deviation and detect outliers. The outlier analysis is performed over all attributes in variety of data sets.

In this current research work, WEKA data mining tool is used to automatically detect outlier. We apply Filters option on unsupervised data and Inter Quartile Range (IQR) on a data set. After applying IQR two attributes are added ,outlier and extreme value .In this paper 7 experiments are performed ,in this some data sets does not have outliers means all instances are normal.

Table 1: Dataset with outlier and extreme

Dataset	Attribute before IQR	Instance	No. of Outlier	No. of Extreme
Diabetes	9	768	49	-
Glass	10	214	16	42
CPU	7	209	36	2
Segmant	20	150	114	424
German_Credit	21	1000	25	153
Weather	5	14	-	-
Iris	150	5	-	-

Table 2: Accuracy using j48 algorithm

Dataset	Accuracy (%)	Accuracy without outlier (%)	Accuracy without extreme (%)
Diabetes	73.82	74.40	74.40
Glass	66.82	68.68	69.93
CPU	96.65	98.48	100
Segmant	95.6	95.23	93.86
German_Credit	70.8	72.82	72.14
Weather	64.28	64.28	64.28
Iris	96.0	96.0	96.0

According to the result from experiment, the accuracies of classifier using dataset with without outliers increase in most datasets. Removing extreme also increases the accuracy in three datasets such as CPU, Glass and Diabetes. The accuracies using other datasets such as Weather, Iris and Segment are not different between before removing outlier and after removing outlier using WEKA tool. So outlier is not excluded in these datasets and it is retaining because it is considered as new trend for further work in data science.

6. Conclusions

The main goal of an outlier detection system is to detect the noisy or abnormal data in the dataset. Furthermore, it is equally important to detect errors at a preprocessing stage in order to reduce their impacts on further classification. This paper presents an outlier detection approach called Box and wisher plot where, the outlier dataset is measured by the J48 algorithm. The trained model consists of seven different datasets. Generally, removing outliers can provide the effective and efficient classification system. However, the experiments are especially sensitive situations when dealing with outliers, omitting an outlier in data can mean omitting information that denotes some new trend or discovery. In this situation, we should not omit outlier, assuming it is not due to an error, it represents a significant success in new discovery. So the outlier should be retained for further new trend research.

7. Acknowledgment

I would like to express my sincere gratitude to Prof. Dr Si Si Mar Win and Prof. Dr Kay Thi Win from the faculty of Computer Studies Mandalay. I also thanks University of Waikato for WEKA tool availability as an open source.

8. References

[1] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

[2] Tsumoto S., (1997)"Automated Discovery of PlausibleRules Based on Rough Sets and Rough Inclusion," Proceedings of the Third Pacific-Asia Conference(PAKDD), Beijing, China, pp 210-219.

[3] Arvind Sharma and P.C. Gupta ―Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ International Journal of Communication and Computer Technologies Volume 01 - No.6, Issue: 02September 2012.

[4] P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool". International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool

[5] R. Delshi Howsalya Devi, M.Indra Devi,A Novel Hybrid Algorithm for Outlier Detection Using WEKA Interface.International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.55 (2015)

[6] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41, 1476–1482

[7] Jerone T. A. Andrews, Edward J. Morton, and Lewis D. Griffin, "Detecting Anomalous Data Using Auto-Encoders," International Journal of Machine Learning and Computing vol.6, no. 1, pp. 21-26, 2016

[8] B. Mantha,"Five Guiding Principles for Realizing the Promise of Big Data," *Business Intelligence Journal*, 2014. [Online] . Available: http://connection.ebscohost.com/c/articles/95066192/five-guiding-principles-realizing-promise-big-data. [Accessed : 04-Mar-2019].

[9] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer Feedback and Data Collection Techniques in Software R&D: A Literature Review," Springer, Cham, 2015, pp. 139–153.

[10] Thomas J. Veasey and Stephen J. Dodson "Anomaly Detection in Application Performance Monitoring Data" *International Journal of Machine Learning and Computing, Vol. 4, No. 2, April 2014*

[11] Sarunya Kanjanawattana, "A Novel Outlier Detection Applied to an Adaptive K-Means," International Journal of Machine Learning and Computing vol. 9, no. 5, pp. 569-574, 2019

[12] Yogitaa, DT. A framework for outlier detection in evolving data streams by weighting attributes in clustering. *2nd IntConf Commun Comput Secur*. 2012;214–222