RESEARCH ARTICLE OPEN ACCESS

Detecting Outliers for Single Dimensional Data Using Interquartile Range

Dr. Kamlesh Malpani
Assistant Professor Department of Computer Science Shri Vaishnav Institute of Management, Indore

ABSTRACT
Outliers are data which can be considered anomalous due to several causes (e.g. erroneous measurements or anomalous process conditions). Outlier detection techniques are used, for instance, to minimize the influence of outliers in the final model to develop, or as a preliminary pre-processing stage before the information conveyed by a signal is elaborated..The traditional outlier detection methods can be classified into four main approaches: distance-based, density-based, clustering-based and distribution-based. Each of these approaches presents advantages and limitations, thus in the recent years many contributions have been proposed to overcome them and improve the quality of the data. In the proposed approach need a specific measurement that will give us an objective standard of what constitutes an outlier. The IQR is used to determine if an extreme value is indeed an outlier. The IQR is based upon part of the five number summary of a data set, namely the first quartile and the third quartile. The calculation of IQR involves a single arithmetic operation. All that to find the IQR is to subtract the first quartile from the third quartile. The resulting difference tells us how spread out the middle half of our data is. Multiplying the IQR by 1.5 will give us a way to determine whether a certain value is an outlier. If subtract 1.5 1× IQR from the first quartile, any data values that are less than this number are considered outliers. Similarly if add 1.5 × IQR to the third quartile, any data values that are greater than this number are considered outliers.

Keywords - Outliers, Detection, Quartile Accuracy ,Efficiency

---

Date Of Submission: 25-08-2019 Date Of Acceptance: 09-09-2019

I. INTRODUCTION

Outlier detection is an important branch in data mining, which is the discovery of data that deviates a lot from other data patterns. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It is used to determine relationships among the internal factors such as price, product positioning, or staff skills, and external factors, such as economic indicators, competition, and customer demographics. Also, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Data mining related methods are often non-parametric, thus, it does not assume an underlying generating model for the data. These methods are designed to manage the large databases from high-dimensional spaces. The identification of outliers can lead to the discovery of unexpected knowledge in areas such as calling card fraud detection, credit card fraud detection, discovering criminal behaviors, computer intrusion detection, etc. Applications such as outlier detection network intrusion detection, customized marketing, weather prediction, pharmaceutical research and exploration in science databases require the detection of outliers. Outlier detection is an important branch in data pre-processing and data mining, as this stage is required in elaboration and mining of data coming from many application fields such as industrial processes, transportation, ecology, public safety, climatology. Outliers are data which can be considered anomalous due to several causes (e.g. erroneous measurements or anomalous process conditions). Outlier detection techniques are used, for instance, to minimize the influence of outliers in the final model to develop, or as a preliminary pre-processing stage before the information conveyed by a signal is elaborated. On the other hand in many applications, such as network intrusion, medical diagnosis or fraud detection, outliers are more interesting than the common samples and outliers detection techniques are used to search for them. The traditional outlier detection methods can be classified into four main approaches: distance-based, density-based, clustering-based and distribution-based. Each of these approaches presents advantages and limitations, thus in the recent years many contributions have been proposed to overcome them and improve the quality of the data. Classical methods are often not suitable to treat some particular databases, therefore recent studies have been conducted on outlier detection for these kind of datasets. In particular, a high number of

contributions based on artificial intelligence, genetic algorithms and image processing have been proposed in order to develop new efficient outliers detection methods that can be suitable in many different applications An exact definition of an outlier often depends on the hidden assumption regarding the data structure and the applied detection method. Some definitions are regarded general enough to cope with the various types of data and methods.

In a sample of moderate size taken from a certain population it appears that one or two values are surprisingly far away from the main group Barnett & Lewis (1978).

An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism Hawkins (1980).

II. CAUSES OF OUTLIERS

Anscombe & Guttman (1960), had attempted to categorize the different ways in which outliers may arise. It was relevant to consider them in rather more detail. In taking observations, different sources of variability can be encountered. We can distinguish three of these. According to Barnett & Lewis (1978)

1 Natural variability

This is the expression of the way in which observations intrinsically vary over the population; such variation is a natural feature of the population and uncontrollable. Thus, for example, the measurements of heights of men will reflect the amount of variability indigenous to that population.

2 Measurement error

Often we must take measurements on members of a population under study. Inadequacies in the measuring instrument superimpose the further degree of variability on the inherent factor. The rounding of obtaining values or mistakes in recording compound the measurement error: they are part of it. Some control of this type of variability is possible.

3 Execution error

A further source of variability arises in the imperfect collection of our data. We may inadvertently choose a biased sample or include individuals who are not truly representative of the population we aimed to sample. Again, sensible precautions may reduce such variability

III. OUTLIER DETECTION TECHNIQUES

Data mining is a useful technique for learning and extracting useful data from a dataset. Various techniques of data mining are designed that helps to search data from the large data set present in the computer. The algorithms that are designed should be highly capable, faster, understandable, robust etc. One of the basic problems of data mining is the outlier detection. An outlier is an observation that is quite dissimilar from the other observations. As various techniques have already been introduced till now, in this section different existing outlier detection techniques have been discussed that are used for detection and removal of outliers.

1 Statistical Outlier Detection

This technique of outlier detection frames the model simply with the help of the data points that are available for processing. In this field most of the outlier research has been done, due to which many distributions of the data is known. Most of the statistical model can only handle one attribute and those that can handle multi attributes can handle data efficiently up to the K<4. This is completely subjective to the distribution used. On the basis of the statistics two methods have been described for the outlier detection. Distribution Outlier Detection
Depth Based Outlier Detection

2 Distance Based Outlier Detection

This is one of the algorithms of outlier detection that is dependent on the distance between the points. The neighbors of the point are selected and checked in this method. If the neighboring points are close then it is considered normal, but if the neighboring point is far away then that case is considered as unusual. This technique is quite efficient as there is no need of defining the explicit distribution that defines the peculiarity.

3 Cluster Based Outlier Detection

This outlier detection technique is quite effective as the data from the datasets is firstly detected without any interference in the clustering process. Various clustering approaches are used for the outlier detection. Clustering on streaming data is categorized by grid based and k means/k median methods. In Partitioning methods, various centroid based methods, k means, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based on RANdomized Search) etc methods are used. One of the clustering is hierarchical clustering. In it, the whole data set is further decomposed into different small datasets. It is further divided into two categories i.e., Agglomerative methods (in which sample units are combined to form single cluster) and divisive

methods (in which single parent cluster is further partitioned)

4 Density based Outlier Detection

In this method of detecting the outlier each object that is present is accredited a LOF (Local Outlier Factor). The local outlier factor is basically the degree assigned to object. In this technique, it is checked that how the object is isolated from its neighbor on the basis of the local outlier. The object with high local outlier factor is termed as outlier and the objects having low local outlier factor are considered to be normal. The high local outlier filter depicts the high probability of being outlier

IV. LITERATURE SURVEY

In 2012 Karanjit Singh and Dr. Shuchita Upadhyaya proposed "Outlier Detection: Applications And Techniques" Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities[1].

In 2013 Jyoti Ranjan proposed "Study of Distance-Based Outlier Detection Methods" An Outlier is an observation which is di_erent from the others in a sample. Usually an anomaly occurs in every data due to measurement error. Anomaly detection is identifying anomalous data for given dataset that does not show normal behavior. Anomaly detection can be classi_ed into three categories: Unsupervised, Supervised and Semi supervised anomaly detection[2].

In 2014 Manish Gupta, Jing Gao proposed "Outlier Detection for Temporal Data: A Survey" In the statistics community, outlier detection for time series data has been studied for decades. Recently, with advances in hardware and software technology, there has been a large body of work on temporal outlier detection from a computational perspective within the computer science community. In particular, advances in hardware technology have enabled the availability of various forms of temporal data collection mechanisms, and advances in software technology have enabled a variety of data management mechanisms[3].

In 2015 Usman Habib, Gerhard Zucker proposed "Outliers Detection Method Using Clustering in Buildings Data" . They discusses the steps involved for detecting outliers in the data obtained from absorption chiller using their On/Off state information. It also proposes a method for automatic detection of On/Off and/or Missing Data status of the chiller. The technique uses two layer K-Means clustering for detecting On/Off as well as Missing Data state of the chiller[4].

In 2016 Kamaljeet Kaur proposed " Comparative Study of Outlier Detection Algorithms" As the dimension of the data is increasing day by day, outlier detection is emerging as one of the active area of research. Finding of the outliers from large data sets is the main problem. Outlier is considered as the pattern that is different from the rest of the patterns present in the data set[5].

In 2017 Rasim M. Alguliyev, Ramiz M. Aliguliyev An Anomaly Detection Based on Optimization "At present, an anomaly detection is one of the important problems in many fields. The rapid growth of data volumes requires the availability of a tool for data processing and analysis of a wide variety of data types. The methods for anomaly detection are designed to detect object's deviations from normal behavior. However, it is difficult to select one tool for all types of anomalies due to the increasing computational complexity and the nature of the data[6].

In 2018 Victoria J. Hodge and Jim Austin proposed "An Evaluation of Classification and Outlier Detection Algorithms "This paper evaluates algorithms for classification and outlier

detection accuracies in temporal data. They focus on algorithms that train and classify rapidly and can be used for systems that need to incorporate new data regularly. Hence, we compare the accuracy of six fast algorithms using a range of well-known time-series datasets[7].

V. PROPOSED APPROACH

The proposed is based on inter quartile range(IQR).In descriptive statistics, the inter quartile range (IQR), also called the mid spread or middle, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data. It is a trimmed estimator, defined as the 25% trimmed range, and is a commonly used robust measure of scale. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

Quartiles are calculated recursively, by using median.

If the number of entries is an even number 2n, then the first quartile Q1 is defined as
First quartile Q1 = median of the n smallest entries
Third quartile Q3 = median of the n largest entries
If the number of entries is an odd number 2n+1, then the first quartile Q1 is defined as
First quartile Q1 = median of the n smallest entries
The third quartile Q3 = median of the n largest entries
The second quartile Q2 is the same as the ordinary median

VI. ALGORITHM OF PROPOSED APPROACH

The proposed approach has the following steps

1. Arrange the given data set into ascending order.
2. Find the median of the given data set by using the formula

If number or even use $\frac{n}{2}$

Or if number or odd use $\frac{n+1}{2}$

Median is denoted as Q2

3. Find the median of upper half and denoted as Q1
4. Find the median of lower half and denoted as Q3
5. Find interquartile range using formula (IQR) = Q3 - Q1
6. Find lower range for outlier by using

$Q_1 - 1.5(IQR)$

7. Find upper range for outlier by using

$Q_3 + 1.5(IQR)$

8. The item which has the value smaller the lower range and has the value greater the upper range are outliers

VII. EXPERIMENTAL ANALYSIS

Proposed approach is compared with two existing approach Mean based method and median based method. We have taken 100 different numeric values to perform the experiment

Table 1 number of data items and missing values

Number of data items	Distance Based approach	Proposed Approach Number of outliers
25	2	4
50	3	5
100	5	9

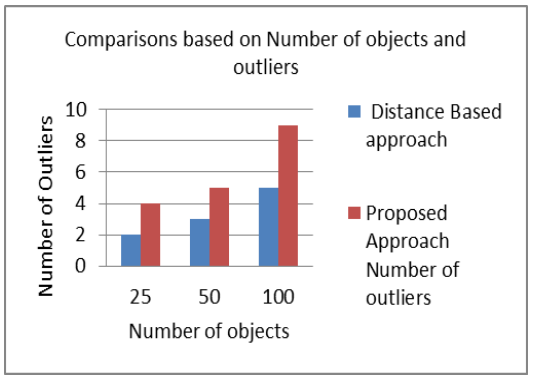

Figure 1 comparison using number of objects and missing value

VIII. CONCLUSION

Outlier detection is an important branch in data mining, which is the discovery of data that deviates a lot from other data patterns. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It is used to determine relationships among the internal factors such as price, product positioning, or staff skills, and external factors, such as economic indicators, competition, and customer demographics. The IQR is used to determine if an extreme value is indeed an outlier. The IQR is based upon part of the five number summary of a data set, namely the first quartile and the third quartile. The calculation of IQR involves a single arithmetic operation. All that to find the IQR is to subtract the first quartile from the third quartile. The resulting difference tells us how spread out the middle half of our data is.

REFERENCES

[1]. Karanjit Singh and Dr. Shuchita Upadhyaya "Outlier Detection: Applications And Techniques" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814 www.IJCSI.org

[2]. Jyoti Ranjan Sethi "Study of Distance-Based Outlier Detection Methods" National Institute Of Technology, Rourkela June 2013

[3]. Manish Gupta, Jing Gao "Outlier Detection for Temporal Data: A Survey" IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 1, January 2014

[4]. Usman Habib, Gerhard Zucker "Outliers Detection Method Using Clustering in Buildings Data" Conference Paper · November 2015 IECON2015-Yokohama November 9-12, 2015

[5]. Kamaljeet Kaur Atul Garg " Comparative Study of Outlier Detection Algorithms" International Journal of Computer

Applications (0975 – 8887) Volume 147 – No. 9, August 2016

[6]. Rasim M. Alguliyev, Ramiz M. Aliguliyev "An Anomaly Detection Based on Optimization" I.J. Intelligent Systems and Applications, 2017, 12, 87-96 Published Online December 2017 in MECS (http://www.mecs-press.org/)DOI: 10.5815/ijisa.2017.12.08

[7]. Victoria J. Hodge and Jim Austin "An Evaluation of Classification and Outlier Detection Algorithms Digital Creativity Labs, Department of Computer Science, University of York, UK {victoria. Hodge, jim.austin}@york.ac.uk 2 May 2018

Dr. Kamlesh Malpni " Detecting Outliers for Single Dimensional Data Using Interquartile Range" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.09, 2019, pp. 31-35