

An Efficient Statistical Based Approach for Outliers Detection

Pragya Soni¹ Mr. Kamlesh Patidar²

¹P.G. Research Scholar ²Assistant Professor

^{1,2}Department of Computer Science and Engineering

^{1,2}JIT Borawan Khargone, India

Abstract— Outliers are suspicious values because they are smaller or larger than the given values. The detection of outliers is important task in interest of data mining that realization that outliers as the key discovery from very large databases. Outlier’s detection is important because of various reasons such as human faults, system error or instrument error etc. Outlier detection techniques are divided into two categories: supervised and unsupervised. Supervised technique assumes the availability of training data set for normal as well as anomaly. Unsupervised technique does not require training data. This approach takes as input a set of data and finds outlier within the data. Sometimes good outliers give useful information for the discovery of new knowledge. Bad outliers are noisy data point. In this paper we propose a novel approach for outlier analysis along with review of some existing outlier detection techniques. In this paper we used some statistical properties to find out outlier. We also find that what the effects are when we consider analysis of data with outlier and without outliers. In this paper we proposed a novel approach in which we used some mathematical relationship between data after deleting the outliers.

Keywords: databases, Machine learning, Data Mining

I. INTRODUCTION

Outlier Detection can be defined as the method of detecting abnormal values which them self-represent uniquely or show them self extremely different from other values in the data set. The aims of outlier detection are find noisy data points. There is no perfect mathematical or biological definition for what constitutes an outlier; ultimately it is a subjective exercise by which we determining whether an observation is an outlier or not. There are several researcher have developed a variety of outlier detection techniques. There is no standardized identification approach they are mainly dependent upon the data set.

Some of the common reasons of outliers include human errors, instrument errors, experimental errors, data processing errors, sampling errors and natural errors Some of the important fields where outlier detection are useful Physics, Economy, Finance, Machine Learning, and Cyber Security[10,11].



Fig. 1: Regression, line with outlier

Consider a data set where scores measured by the sales in which two outliers are shown with blue color points

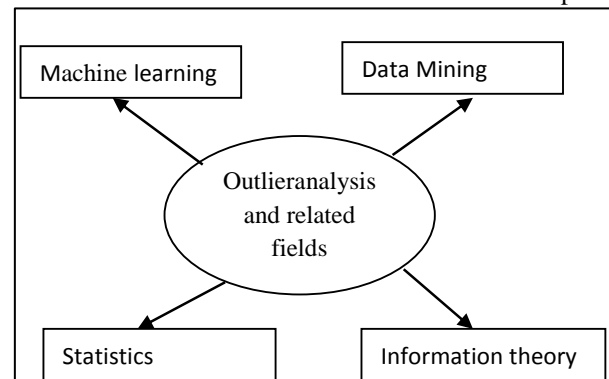


Fig. 2: Other field related with outlier analysis

II. METHODS FOR OUTLIER DETECTION

- Extreme Value Analysis
- Statistical Modeling
- Proximity Based Models
- Information Theory Models
- High Dimensional Outlier Detection Methods

A. Extreme Value Analysis

Extreme Value Analysis is basic form of outlier detection for 1-dimension data. This approach assumed that values which are very large or very small are outliers. Z-test and t-test are some common examples. This approach is good for primary analysis of data but they do not have better result for high dimensional value. Extreme Value Analysis are mostly used for interpreting outputs of other outlier detection methods [12, 13].

B. Statistical Models

This approach, divides the data into a lower-dimensional space with the help of linear correlations. I the next step this approach find distance of each data point to a plane after fitting the data into sub-space. Now the calculated distance is used to discover outliers. Principal Component Analysis is very common example of Statistical Models for outlier detection.

C. Proximity-based Models

This approach is demonstrated isolated point from the rest of the values. Some of the common examples includes cluster analysis, density-based analysis, and nearest neighborhood. The basic concepts in proximity-based methods are isolated outliers from the remaining data. This approach used three basic clustering approaches cluster analysis, density-based analysis and nearest neighbor analysis.

D. Information-Theoretic Models

This approach used minimum code to describe outliers for a given data set. This method provides abstract summary of the given data points, the deviations from which are labeled as

outliers. Information theoretic measures are broadly based on this principle. The key concept behind this approach is that outliers the minimum code length required to describe a data set.

E. High Dimensional Outlier Detection Methods

In many real life applications, data sets contain several features. The traditional approaches for outlier detection such as PCA and LOF will not be able to handle and effective. Density-based Outlier methods are effective method to find outliers in high dimensional data sets. LOF method is used for the data set to calculate the nearest neighborhood of each data point and finally outlier score for each data point.

III. LITERATURE SURVEY

In 2014 Kamal Malik and H.Sadawarti proposed "Comparative Analysis of Outlier Detection Techniques". They provide the broad and comprehensive literature survey of outliers and outlier detection techniques under one roof. They explained the complexity associated with each outlier detection technique. They also gave a broad comparison of the various methods of the different outlier techniques. They emphasized that there is no universally accepted gamut of any methodology to detect and analyze the outliers. [3].

In 2015 Shivani P. Patel and Vinita Shah proposed "A Survey of Outlier Detection in Data Mining". Outlier detection plays an important role in data mining. They showed that outlier detections are useful in fields like network intrusion detection, credit card fraud detection, stock market analysis, detecting outlying in wireless sensor network data, etc. They gave a brief survey on different outlier detection approaches, which are statistical-based approach, deviation-based approach, distance-based approach, density-based approach. They also gave comparison in different approaches of outlier detection which are statistical approach, distance-based approach, density-based approach, deviation-based approach. [4].

In 2016 Kamaljeet Kaur and Atul Garg proposed "Comparative Study of Outlier Detection algorithms". They covered various outlier detection algorithms like statistical based outlier detection, depth based outlier detection, clustering based technique, density based outlier detection etc. They compared study of outlier detection methods and find out which of the outlier detection algorithms are more applicable for high dimensional data points. They presented the study of different existing outlier detection techniques and the way in which they are categorized. They also concluded performance of clustering algorithms is comparatively better than other outlier detection algorithms on huge data sets [5].

In 2017 Dipannita Kar, Mr. Haresh Chande and Mr. Rajendra Gaikwad proposed "A Study Paper on Outlier Detection on Time Series Data". They proposed a new approach to detect the outliers. In the proposed algorithm time taken as an important attribute of each dataset. They showed that it is important for each process of data mining to give more accurate and useful information. They also analyzed with the help of WEKA tool. They also worked on K-Mean, Density based, EM, and Cobweb. They gave comparative study has been performed on the k-means, Density based, EM and Cobweb algorithm [6].

In 2018 Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gaze proposed "ICS Outlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure". They proposed a new approach Invariant Coordinate Selection (ICS). They showed remarkable properties for identifying outliers in low-dimensional subspace with invariant components. They also implemented in the ICS Outlier package. They compared proposed approach with approaches and found that ICS is efficient give advantage in the context of a small proportion of outliers [7].

In 2019 Paulo Joao and Octavian Postolache proposed "Healthcare Outlier Detection with Hierarchical Self-Organizing Map". They proposed the use of Hierarchical Self-Organizing Map (HSOM) algorithm to perform clustering analysis. They used dimensionality reduction for outlier detection in healthcare data. They provided an appropriate framework and performed clustered task based on individual data. This included the standardization and enhances the benefits for various applications. They also solved the problem of the cluster border effect and produces partial clusters for the edge of the U-mat. [8].

In 2020 Harry Bhagat, S.Priya and K. Aditya proposed "Outlier Detection Based on Machine Learning Techniques". They presented better understanding of the different approaches of research on outlier detection. They came out with perfect results to find anomaly using various approach and decrease the fraud pricing of housing. They showed that proposed work made the easy and find the price whether its suitable for that society. They showed the significance of information; they can convert into noteworthy data in a wide assortment of utilizations. A strange traffic design in a PC system could imply that a hacked PC is conveying touchy information to an unapproved goal [9].

IV. PROBLEM STATEMENT

- 1) Outliers are observed data points that are far from the least squares line.
- 2) These are "errors", and can be residual (vertical distance) from the line to the point.
- 3) Outliers may have a big effect on the slope of the regression line. To begin to identify an influential point, how we can remove it from the data set and see if the slope of the regression line is changed significantly. How we could guess at outliers by looking at a graph of the scatter plot and best fit-line. We would need some guideline as to how far away a point needs to be in order to be considered an outlier. How to find new line which is a better fit to the remaining data values. The line can better predict the final exam score given the third exam score.

V. PROPOSED ALGORITHM

Proposed algorithm has following steps

- 1) Draw the scatterplot. Linear or non-linear pattern of the data Deviations from the pattern outliers.
- 2) Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot and normal probability plot (for normality assumption). If the assumptions of the model appear not to be met, a transformation may be necessary.

- 3) Fit the least-squares regression line data by calculating these values
- 4) Calculate SSE Sum of Squared Errors and calculate correlation coefficient. Calculate s , the standard deviation
- 5) Measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $2s$.
- 6) The data point which do not satisfy the condition declared as outliers

VI. IMPLEMENTATION AND RESULT ANALYSIS

No of Objects	Sum of Squared Errors with outlier	Sum of Squared Errors without outlier
100	2440	108
200	2456	106
300	2780	123

Table 1: Comparisons using SSE with outlier and without outlier

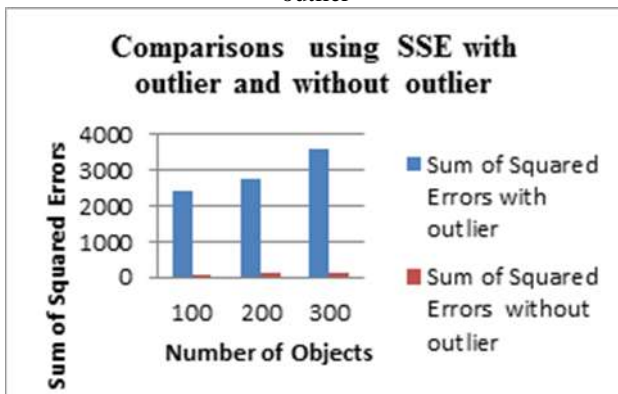


Fig. 2: Comparisons of SSE with and without outlier

VII. CONCLUSION AND FUTURE WORK

There are several algorithms and methods have been developed for classification. The most popular classification methods are artificial neural networks, Decision Tree, Support Vector Machine Naïve Bayes Classifier and K-Nearest Neighbor Classifier. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. In the proposed work we used K-Nearest Neighbor Classifier Algorithm and analysis this algorithm for different value of K. From experiment it clear that value of should always take odd number. The value of parameter K is taken in such a way so that it is not very small and not very big. Deciding value of k is a continue process until number of records are correctly classified. K-NN Algorithm is also simple to understand and calculation is easy.

REFERENCES

- [1] Lakshmi Sreenivasa Reddy.D DrB.RaveendraBabu Outlier Analysis of Categorical Data using FuzzyAVF2013 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2013]
- [2] VarunChandolaArindam Banerjee Outlier Detection : A Survey Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications| Data

- Mining General Terms: Algorithms Additional Key Words and Phrases: Outlier Detection, Anomaly Detection
- [3] Kamal Malik1 H.Sadawarti2, Member IEEE, 3Kalra G.S., Member IEEE Comparative Analysis of Outlier Detection Techniques International Journal of Computer Applications (0975 – 8887) Volume 97– No.8, July 2014
- [4] Shivani P. Patel Vinita Shah A Survey Of Outlier Detection In Data Mining National Conference on Recent Research in Engineering and Technology (NCRRET -2015) International Journal of Advance Engineering and Research Development (IJAERD) e-ISSN: 2348 - 4470 , print-ISSN:2348-6406
- [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] DipannitaKar, Mr. HareshChande, Mr. RajendraGaikwadA Study Paper on Outlier Detection on Time Series Data www.ijcrt.org © 2017 IJCRT | Volume 5, Issue 4 December 2017 | ISSN: 2320-2882
- [7] Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure The R Journal Vol. 10/1, July 2018 ISSN 2073-4859
- [8] Paulo Joao Octavian Postolache Healthcare Outlier Detection with Hierarchical Self-Organizing Map ©2019 IEEE Healthcare Outlier Detection with Hierarchical Self-Organizing Map August 2019 DOI: 10.1109/ISSI47111.2019.9043675.
- [9] Harry Bhagat1, *S.Priya2, K. Aditya3 Outlier Detection Based on Machine Learning Techniques International Journal of Advanced Science and Technology Vol. 29, No. 6, (2020), pp. 2142 - 2151
- [10] Tung Kieu , Bin Yang_ , ChenjuanGuo and Christian S. Jensen Outlier Detection for Time Series with Recurrent Auto encoder Ensembles Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).
- [11] Stefan Mandić-Rajčević * and Claudio Colosio A Proposed Approach and Validation Using Biological Monitoring Department of Health Sciences, University of Milan and International * Correspondence: stefan.mandic-rajcevic@unimi.it Received: 20 June 2019; Accepted: 9 July 2019; Published: 12 July 2019
- [12] ZeeshanAhmad Lodhia1and Akhtar Rasool2, and GauravHajela A survey on machine learning and outlier detection techniques IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, May 2017
- [13] C. Leela Krishna1*, C. Kala Krishna2 Outlier Detection Using Association Rule Mining and Cluster Analysis International Journal of Computer Sciences and Engineering Open Access Research Paper Vol.-6, Issue-6, Jun 2018 E-ISSN: 2347-2693