ISSN 2229-5518

To Detect Outlier for Categorical Data Streaming

MANOJ MISHRA[1], NITESH GUPTA[2]

ABSTRACT-Instant identification of outlier patterns is very important in modern-day engineering problems such as credit card fraud detection and network intrusion detection. Most previous studies focused on finding outliers that are hidden in numerical datasets. Unfortunately, those outlier detection methods were not directly applicable to real life transaction databases. Outlier detection methods are divided into transaction specific and non transaction specific outlier detection methods, in this paper we are going to focus mainly on transaction specific methods and detect outlier transactions from transactional databases e.g. purchase of the data at the store, customer dataset at a company. Here we are going to compare two transaction specific methods and find efficient method from them.

KEYWORDS- Outlier, Categorical Data Streaming, and Transaction based method, association rule, and frequent pattern.

I. INTRODUCTION-

In a Database, outlier is remaining objects which are remained in mining of data. Outlier detection is the process to detecting dataset or instances which are arise in a system for its unusual behaviour. Effective detection of outliers can moderate to the discovery of valuable messages in the data. After many years, mining for outliers has received significant attention due to its wide applicability in areas such as detecting fraudulent of credit cards, intrusion detection system, and unauthorized access in computer networks, medical diagnosis weather prediction and environmental monitoring. A number of existing methods are designed for detecting outliers in continuous data. Most of these methods employ distances between data points to detect outliers. In the case of data with categorical attributes, undertakes are often made to map categorical features to numerical values. Such mappings enforce arbitrary ordering of categorical values and may cause unreliable result. Another problem is related to the big data phenomenon.

- *Manoj Mishra, MTECH (CSE), NIIST, Bhopal. Affilied to RGPV, Bhopal, M.P, India er.mnjms@gmail.com*
- *NITESH GUPTA, Asst. Professor (CSE) in NIIST, Bhopal. Affiliated to RGPV Bhopal, M.P, India. 9.nitesh@gmail.com*

Today's, Many systems are able to generate and capture real-time data continuously. Some examples include real-time data acquisition systems, condition monitoring systems, and sales transaction systems. It is a challenging task to effectively detect outliers occurring in categorical data streams. Traditional outlier detection approaches are no longer feasible as they only deal with statics data sets and require multiple scans of data to produce effective results. In data streams setting, outlier detection algorithms (e.g., [8]) need to process each data item within a strict time constraint and can only afford to analyze the entire data set with a single scan of data.

In outliers detection has widely research problem in credit card fraud detection, fault-tolerance, anomaly detection, medical diagnosis in categorical data streaming. Outliers are non-conforming patterns in data; that is, they are patterns that do not exhibit normal behaviour.

There are two outlier detection methods transaction specific and non transaction specific outlier detection methods. We are going to focus on transaction specific outlier detection methods. In this paper we are going to compare two methods association rule based outlier detection method and frequent pattern based outlier detection method.

II. OUTLIER DETECTION METHOD:

Statistical Outlier Detection: Statistical method of construct data distribution model. Based on model it declares point as outlier. But it observed that it is

applicable for single dimension. Parametric assumption also does not hold well on distributional data set.

Depth based Outlier Detection: This method is belongs to statistical outlier detection but it is independent of data distribution.

Distance based Outlier Detection: This approach use a notion of distance of data point within data distribution and by using threshold value it decides outlier within data .It suffers from detecting a local outlier within multi categorical data or diverse density data. It suffers from high computational cost for high dimensional data set.

Density based Outlier Detection: This approach uses a density estimation of data and the data element which has allow density is declared as outlier. Selecting boundaries for outlier is difficult task. We require lower bound and upper bound for selection of parameter.

Cluster based Outlier Detection: This approach uses a cluster based technique for detecting outlier where it finds closely related objects. Object which does not belong to any cluster or belongs to a small cluster is declared as outlier. A major limitation of clustering-based approaches to outlier detection is that they require multiple passes to process the data set. Outlier detection also highly depends upon type of clustering used.

III. LITERATURE REVIEW-

Outlier detection methods depend on whether the data are quantitative or categorical. A large number of books and articles are devoted to the detection of outliers in quantitative data (See, for example, [2], [3], [10], [11], [12], [13], and [14]). Although categorical data are common in many diverse domains such as business, medicine, psychology, sociology, education, and political science [5] by contrast with quantitative data, a relatively few articles are devoted to the identification of outliers in categorical data (See, [4], [5], [6], [7], [8], [9], [15]). The reason for this imbalance in the literature coverage is that the problem of identification of outliers in categorical data is much more challenging than the problem of identifying outliers in quantitative data.

At. Ayman Taha and Ali S. Hadi , Outliers identification algorithms for categorical datasets strongly depend on parameter settings that require prior information about the data, e.g. number of outliers in the data, maximum length of item sets and/or minimum support for frequent item sets. These input parameters are classified into two groups; (a) intrinsic parameters which are required by an outlier's detection method to produce a score measure to each object and (b) decision parameters which are required for deciding on whether an object is an outlier based on the score. In this paper, a general approach for automating decision parameters of outlier's identification in multivariate categorical data is proposed. The added value of the proposed approach is that it can be used by any outlier's detection algorithm for categorical data that produces a score measure for each object. We provide a simulation approach for computing critical values for any outlier's detection algorithm. These critical values are distribution-free statistical measures. They are also based on data-driven characteristics; hence they can be used for the identification of outliers based on the score measure produced by the algorithm. We illustrate this approach using two outlier's detection algorithms. Furthermore, real and synthetic datasets are used to evaluate the performance of the proposed approach. In this paper, we pointed out that the available techniques for outlier identification in categorical datasets require the user to specify many decision parameters such as the number of outliers in the data. We propose a general approach to compute robust statistical critical values based on data driven measures. Algorithms for generating categorical datasets with and without outliers are provided. These data are used to compute critical values and also to test the performance of the proposed approach. We illustrate the approach using two outlier detection methods, but the approach is quite general and is applicable to many outliers' identification methods. The intrinsic idea of the proposed approach is employing data driven measures to calculate these statistical critical values. We proposed and investigated three alternative distribution free methods for the automatically identification of outliers based on regression model. Real and synthetic datasets are used to evaluate performance of the approach.

In statistical approach the parameters are computed assuming all data points have been generated by a statistical distribution like Gaussian method. Outliers are points that have a low probability to be generated by the overall distribution. In depth based approach outliers are located at the border of the data space. Normal objects are in the centre of the data space. In distance based approach [17] normal

data objects have a dense neighborhood.Outliers are far apart from their neighbours, i.e., have a less dense neighbourhood. In density based approach an outlier is considerably different to the density around its neighbours i.e. outlier score. Classification [16] and clustering [18] are also non transaction specific methods.

IV.PROBLEM FORMULATION-

Mining useful and interesting information from a large amount of data is prior task in data analysis. Designing an appropriate outlier detection technique is more challenging with respective context of streaming data and nature of data. With reference to following challenges traditional outlier detection methods might be not able to detect outliers over streaming data.

□*Distributed Streaming Data.*
Data coming from distributed environment may dynamically change. In such situation distribution of streaming data may not known. Due to dynamic nature direct computation of probability is difficult. Traditional methods are offline in manner do not able to handle distributed streaming data.

□*Massive Data Processing.*
Data stream are having massive amount of data. Due to property of dynamic change of data it difficult to process data within single scan. In such situation prior challenge to traditional outlier detection is to provide a high detection rate. It is observed that traditional outlier detection techniques do not scale well to process large amount of distributed data streams in an online manner.

□*Limited Computation Resources.*
In many application domain, high computation power and the other measuring property such consumption of available memory at hand does not measure up the massive amount of streaming data. For example in single day Google provide search for 150 million query search, Telstra generated 15 million call records. Traditional outlier detection methods are not able to handle such requirement for streaming data. Stream mining algorithms shall learn fast and consume little memory resources.

□*Uncertain and Missing Data.*
In most application domain that we don't have a sufficient data for operations. Such situation arouse due to uncertain and missing data. If we don't have sufficient information for data that may lead us for wrong decision. We required a method to manage such uncertain data and missing data within data stream.

□*High Dimensional Data Stream.*
High dimensional data stream contains a tremendous huge amount of data. Such massive amount data contains a large data with high dimensions with data complexity. For example wireless sensor network data, web logs, Google search, etc. Traditional methods are not suitable over high dimensional data as they required very high computation cost for processing data.

Abstractly speaking outliers are patterns that deviate from expected normal behaviour, which in its simplest form could be represented by a region and visualize all normal observations to belong to this normal region and consider the rest as outliers This approach looks simple but is highly challenging due to following reasons.
It is very difficult to define the normal behaviour or a normal region. The difficulties are as under.
□encompassing of every possible normal behaviour in the region.
□imprecise boundary between normal and outlier behaviour since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.
□Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from malicious actions.
□In many domains normal behavior keeps evolving and may not be current to be a representative in the future.
□Differing notion of outliers in different application domains makes it difficult to apply
technique developed in one domain to another.
For example, in the medical domain a small deviation from normal body temperature might
be an outlier, while similar value deviation in the stock market domain might be considered as normal. Even within same domain say crime detection there could be situations where use of foreign make weapons may be considered normal in crimes committed in metro cities but an outlier for murders of commoners in tribal regions.
□Availability of labeled data for training/ validation of models used by outlier detection
techniques.
□Noise in the data which tends to be similar to the actual outliers and hence difficult to distinguish and remove.
Due to the above challenges, the outlier detection problem, in its most general form, is not easy to solve. In fact, most of the existing outlier detection

techniques solve a specific problem formulation which is induced by various factors such as nature of the data, availability of labeled data, type of outliers to be detected, etc. Often, these factors are determined by the application domain in which the outliers need to be detected.

V. PROPOSED WORK-

Our proposed approach for categorical data stream to investigation of critical value sensitive for better performance measures. We proposed the Transaction based method for categorical data streaming. Transaction specific outlier detection methods are association rule based outlier detection method and frequent pattern outlier detection method.

They have specified that they get scores from the different value, collect them and then evaluate the outlier based on all the scores.

We can have various examples here.
1) As suppose in stock market a company value will be depending on various factors and ultimately outlier will be obtaining from the scores of those multiple factors.

Stock market companies- CSi;

Factors of each Company – Fi;

Outliers based for user U is based on the all the average and then thresholding from them.

Peak User value=Pu.

For Each Factor (user=1; user<Ui; user++)

PUV= P (user)>=Ti;

For (user=1; user<Ui; user++)

Outlier=PUV>=Ti;

As in stock market Stock price Stock purchased Stock price variation all will generate score and based on all scores a outlier can fall to take decision in any case.

2) In ranking case for any scenario such as we want to provide any sort of certification or affiliation to any business will be given by scoring multiple factors not will be with single factor.

As suppose for university-Campus Area Facility will be playing crucial role to find outlier if we take multiple university and their affiliation question.

VI. CONCLUSION-

In this paper, we target transaction databases and propose the detection of transactions that are likely to be outliers. Defining the concept of associative closure using association rules with high confidence, we derive a formula for outlier degree. Thus from these paper we conclude that the transaction specific method Association rule based outlier detection method is more efficient than frequent pattern outlier detection method using FPOF score for outlier transaction detection.
Speed of FindFPOF method is slow as first for discovering the frequent patterns in the data set. FindFPOF method uses the Apriori algorithm as algorithm for finding the frequent patterns, which is time-consuming. Association rule based method is having higher accuracy than FindFPOF method as shown from the experimental results. Since infrequent items will induce incorrect calculation on outlier degrees, the proposed method modified the definition of transaction's association closure by removing the infrequent items before the calculation of outlier degrees. The experimental results provide evidences to verify that the proposed algorithm is more efficient in both accuracy and precision rates. Thus association rule based method is more efficient than frequent pattern based outlier detection method.

VII. REFERENCES-

1. Ayman Taha and Ali S. Hadi, "A General Approach for Automating Outliers Identification in Categorical Data", 978-1-4799-0792-2/13/$31.00 ©2013 IEEE.
2. S. Jiang, X. Song, H. Wang, J.-J. Han, and Q.-H. Li, "A clustering-based method for unsupervised intrusion detections," *Pattern Recognition Letters*, vol. 27, pp. 802–810, 2006.
3. H. Cheng, P.-N. Tan, C. Potter, and S. A. Klooster, "Detection and characterization of anomalies in multivariate time series," in *Proceedings of the SIAM (SDM)*, pp. 413– 424, 2009.
4. A. Koufakou, E. Ortiz, M. Georgiopoulos, G. Anagnostopoulos, and K. Reynolds, "A scalable and efficient outlier detection strategy for categorical

data," in *Proceedings of the IEEE ICTAI*, (Patras-Peloponnese- Greece), 29–31 October 2007.
5. A. Taha and O. Hegazy, "A proposed outliers identification algorithm for categorical data sets," in *Proceedings of the INFOS*, (Cairo, Egypt), pp. 1–5, 2010.
6. Z. He, X. Xu, and S. Deng, "A fast greedy algorithm for outlier mining," in *Proceedings of the PAKDD*, pp. 567–576, 2006.
7. S. Li, R. Lee, and S.-D. Lang, "Mining distance-based outliers from categorical data," in *Proceedings of the IEEE ICDM Workshops*, pp. 225– 230, 2007.
8. K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *Proceedings of the ACM KDD*, pp. 220–229, 2007.
9. K. Narita and H. Kitagawa, "Detecting outliers in categorical record databases based on attribute associations," *Progress in WWW Research and Development*, vol. 4976, pp. 111 123, 2008.
10. N. Billor, A. S. Hadi, and P. Velleman, "Blocked adaptive computationally-efficient outlier nominators," *Computational Statistics and Data Analysis*, vol. 34, pp. 279–298, 2000.
11. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41(3), pp. 1–58, 2009.
12. H.-P. Kriegel, P. Kr¨oger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in *Proceedings of IEEE ICDM*, 2012.
13. E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, "On evaluation of outlier rankings and outlier scores," in *Proceedings of the SIAM (SDM)*, pp. 1047–1058, 2012.
14. C. B¨ohm, K. Haegler, N. M¨uller, and C.Plant, "Coco parameter-free outlier detection with coding costs," in *Proceedings of the ACM KDD*, pp. 149–158, 2009.
15. K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in *Proceedings of the ACM KDD*, pp. 169–176, 2008.
16.Dr. Shuchita Upadhyaya, Karanjit Singh, "Classification based outlier detection techniques"
17. Knorr E.M., Ng R.T., Tucakov V., "Distance based method: algorithm and applications"
18. Rajendra Pamula,"Outlier detection method based on clustering"