IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. I (Jan - Feb 2020), PP 01-07
www.iosrjournals.org

Outlier Detection using different clustering Approaches

Harshada Mandhare[1], Mohammad Shahid[2]
[1]*Assistant ProfessorDept. of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur, Maharashtra.*
[2]*Assistant ProfessorDept. of Computer Science and Engineering,G H Raisoni College of Engineering,Nagpur, Maharashatra.*

Abstract: *Recognition of objects, measures or remarks that doesn't match to a predictable sample or further objects in a different record set is called as outlier detection. Outlier detection is valid in different areas such as data analysis, disease prediction, fraud detection, organization fitness checking, and result findings in antenna systems, plus discovering Eco-system disorders. It is mainly utilized in data processing to eliminate the abnormal record from the different record set. The progress in special statistical expertise, the different number of data, in addition to their dimension and complexity which raise vastly, therefore it show result in the requirement of programmed inspection in the massive capacity of assorted structured record. So the plan for various structured record is nothing but special data mining techniques are operated. The goals of the different types of schemes or methods are to identify invisible reliance from their different group of records. In this paper we will propose three special outlier detection techniques such as Cluster based outlier detection, Distance based outlier detection and Density based outlier detection. We will use different three datasets related to health care such as melonama, esophageal cancer, and Pima dataset to calculate the numbers of outliers from these datasets for accurate data retrieval. Also will perform the comparison between these three techniques to ensure which technique offers better performance.*

Keywords: *outlier, data mining, distance based outlier detection, K-means clustering, density based clustering, cobweb algorithm.*

Date of Submission: 24-12-2019 Date of Acceptance: 07-01-2020

I. Introduction

Outliers' is defined as observations or remarks in group of records that emerge to be incoherent amid the remains of that group of record, or it turn so greatly since additional observations or remarks so as to produce doubts that they were produced via a special method. The recognition of outliers or remarks those are able to direct the finding of helpful information as well as number of realistic appliances in the fields like credit or debit card deception discovery, competitor presentation investigation, election abnormality investigation, plus strict climate calculation. Unofficially, a spatial outlier is nothing but limited insecurity, or a severe remark with their nearby rates, still although it might not be extensively special since the complete populace. Discovering spatial types of outliers or remarks is helpful in several appliances of geographical record organizations, containing shipping, biology, community security, community fitness, agrology, as well as position related methods [1]. Conventional outline acknowledgment intends to discover the universal outline for their mainstream of record plus extravagances outliers as error. This might outcome in defeat of vital unseen record as single human being error might be a further human being indication. There are numerous areas, outliers are vital issue than the ordinary record, because they might show whichever abnormal activities or else starting of fresh outline, that can reason of injure to the customer. The outlier discovery goal is to discovering the unusual record who's activities is extremely exclusion evaluated through extra record, this turn into increasingly helpful device in various appliances, like bank card deception discovery, remedy investigate, finance sanction, interruption discovery, advertising with purchaser distribution etc. Mainly schemes of outlier record removal in recent implementation are totally related to information. These types of schemes are essentially categorized in to couple of parts: division and depth related scheme. In a partition schemes it assists regular division to fit the proper record. Outliers are described with respect to possibility division. Division related scheme difficulty is that it imagines that the essential record division is recognized a proceeding. Still, for numerous appliances, the previous information is inconstantly accessible, as well as the rate for welling record through regular division is considerably significant [2]. In distance Based detection scheme, the easiest plus mainly usually utilized methods, since it simply analyze the space among the different entities. Also distance is an extensively utilized set in numerous record removal difficulties. Several appliances like bank card deception plus economic investigate in the entire procedure large dimensional records. Mainly it is complex to discover outliers in these gaps openly, since it is very solid to visualize the division of records positions in large dimensional gaps. In

detail, statistics turn into spare in large dimensional gaps, therefore it is very complex to differentiate entities through the calculation of compactness. As a result, distance related schemes misplace their implication while selling through troubles that practice large dimensional records [3].

Enhanced density related space compute is projected to be utilized by a progressive exploration algorithm for the outlier discovering. The compactness related space compute is shown to decrease the different numeral of couple calculations required to recognize outlier. A new rearrangement operative which develops the compactness based record is too projected through the progressive algorithm [4]. A dispersed technique is described for discovering distance related outliers in huge records. The technique is stands on the idea of outlier discovery explaining group that is tiny sub record of the record set which might be too engaged for calculating new outliers. The technique develops equivalent calculation in categorize to gain huge instance accumulating. Certainly, past defending the accuracy of the outcome, the projected scheme show brilliant presentations. Since the academic position of inspection, for universal locations, the sequential rate of algorithm is predicated that the three commands of size quicker than the standard nested round such as method to discover outliers [5]. The idea of the distance related outliers exclusively, the outlier discovery could be completed expertly for the large group of records, in addition to for the k measurement group of records during the large charges of k. Secondly outlier discovery is a significant with significant facts finding job [6]. The method for the distance related outliers is to utilize the region assets of the trouble to division calculation between the cores of a multi-core structure or the main join of an announcement group that to acquire huge instance economies [7]. In the partition related outlier discovering algorithm, primary it divide the input positions with a clustering algorithm, through it calculates lesser as well as higher vaults for positions in the every division. It occupies these types of group of record to recognize the divisions which does not probably control the peak outliers as well as reduces them. Outliers are calculated since the continuing positions in an ending stage. Since percentage is normally tiny, algorithm reduces an important amount of positions, in addition the outcomes in extensive economies in the total of calculation [8]. The partially, gap position intensity of a position is a record set, as the minimum numeral of observations or remarks in several blocked partially gap by limit. Since partially gap intensity might be a type of multi-various grading. The earnest position with maximum partially gap intensity, is a multi-various simplification of the middle. At this time, the earnest position might simply be calculated for over-rate record. An algorithm called deep location is used to estimate the earnest position in the large dimensions [9]. To utilize numerous data theory procedures, explicitly, decline, qualified decline, comparative qualified decline, record expands, plus record rate for abnormality discovery. These types of evaluates might be utilized to explain the character of an review record set, recommend the suitable abnormality discovery form to be constructed, as well as clarify the presentation of the form [10]. In this paper we will implement novel different outlier detection techniques such as, Cluster based outlier, and Distance based outlier plus Density based outlier detection technique. For execution, we will use three dissimilar dataset to determine the outliers and will also show comparative analysis between these three algorithms that will show the which outlier detection technique provides superior correctness than other two outlier detection technique. This paper will gives special aspects such as it will help to evaluate three outlier detection techniques with the help different parameters that are no. of clusters, outliers and execution time for determining outliers and clusters. Also it will show comparison between three different techniques that will show which technique offers improved effectiveness and correctness.

This paper is organized in following section: Section 3 explored the existing implementation related to the outliers' detection methods. Section 4 proposes novel three different techniques which help to determine the outliers to recover perfect statistics. In section 5 we will define the system architecture. Lastly in section 6 we draw a conclusion and future scope in data mining system area.

II. Background And Motivation

In data mining, outlier finding is considered as a difficulty of discovering samples in the record set that not able to get predictable standard activities. These types of inconsistent samples are frequently assumed as an outliers, abnormalities, inharmonious remarks, omissions, errors, imperfections, deviations, harm, revelation, innovation, customs or else impurity in special appliance fields. Outlier finding have been an extensively investigated difficulty plus discovers enormous exploit in an extensive range of appliances fields like banking area, cover, income tax deception finding, incursion finding for pretend protection, liability finding in protection vital organizations, forces inspection for the rival behaviors as well as several additional fields. Significance of outlier discovery is appropriate information that the outliers in record set transform to considerable information

in an extensive range of appliance fields. Outlier finding have been searched to be straightforwardly valid in a huge numeral of fields that consequence in a massive as well as extremely different studies of outlier finding methods. These different types of methods have been expanded to resolved centered difficulties affecting to the different appliance field, whereas furthers has been implemented in the supplementary general approach.

In outlier discovery a vital challenge is nothing but it occupies searching the hidden gap and showing a standard area that includes all feasible standard activities is extremely complex. In numerous studies, outliers or remarks are the outcome of nasty measures; the nasty challengers adjust individually to build the distant observations or remarks occur such as regular, thus building the job of essential regular activities added complex etc. Hence to decrease the diverse type of extortion difficulties, outlier detection techniques are important to find the outliers from the different records to access exact and essential record.

III. Existing System

Bo Liu et.al. has presented a new outlier finding technique to discover record among inadequate tags also integrate the inadequate irregular tags into knowledge [11]. To contract with the record through inadequate tags, they have established possibility rates for every input record that helps to signify the quantity of link of an instance to the regular as well as irregular groups correspondingly. In this they have worked in two different stages. In first step, they have generated a simulated exercise record set through calculating possibility rates of every pattern related to its limited activities. Also they have presented kernel *k*-means clustering algorithm as well as kernel local outlier related technique that help to calculate the possibility rates. In second stage, they integrated the produced possibility rates as well as restricted irregular instances into other knowledge structure that construct a new correct category for worldwide outlier discovery. For incorporating restricted as well as worldwide outlier discovery, projected technique openly holds record through defective patterns plus improves the presentation of outlier discovery. Markus M. Breunig et. al. have considered numerous of different situations, that is supplementary significant to allocate every entity a grade of individual an outlier or remark. This grade is known as the restricted outlier aspect of an entity [12]. It is restricted that the grade rely on how secluded the entity is through the contiguous region. A complete prescribed investigation shown that the local outlier factor has several admired assets. With the original record set, they have verified that this might be utilized to discover outliers that show very significant. Lastly, a suspicious presentation estimation of proposed algorithm proved that proposed technique of discovering restricted outliers might be convenient. Rashi Bansal et. al. has defined many different outlier discovery appliances as well as methods in data mining area [13]. Outlier discovery have extreme accomplishment exploit in huge variety of interruption discovery, cellular phone plus assurance maintain deception discovery, medicinal as well as community fitness outlier discovery in addition to manufacturing harm discovery. As there are vast records of techniques that are used to absolute this kind of job, along with normally group of the largely correct technique causes a vast test to the professional. Fabrizio Angiulli et. al. has proposed new technique ,called Dolphin, which is used to discover distance related outliers in huge dimensional data [14]. The projected technique executes on couple of in order inspects of the record set that are requires to accumulate into core memory location as a main fraction of the record set, to competently explore for nearer as well as early on reduce infers. The approach practiced via this technique that permits to keep that fraction extremely tiny. Equally academic validations as well as experiential proof that the dimension of the accumulated total record to a small number of percentages of the record set that are offered. An additional significant characteristic of technique is that the memory neighborhood record is listed via an appropriate near checking method. This allows searching for nearest neighbors searching simply at a tiny subgroup of the core memory accumulated record. Sequential plus dimensional rate examination has presented that this technique accomplished couple of continuous processor along with I/O charge. This technique has evaluated during the condition of the skill schemes that show it exceeds offered schemes.

Elio Lozano et.al. has implemented couple of equivalent methods; primarily method is used to discover distance related outliers that are rely on layered rounds via inconsideration with the utilize of a reducing law [15]. The secondarily method is used to discover compactness related restricted outliers or remarks. In this two methods record parallelism is utilized. They have shown that the both methods attain by continuous accelerate. These two methods have done testing on four different original record groups approaching from the Machine discovering storage warehouse. Charu C. Aggarwal et.al. have present different types of methods for outlier discovery that discover the outliers via learning the activities of ledges from the record [16]. Mohiuddin Ahmed and Abdun Naser Mahmood have investigated an innovative unverified method to recognize outliers or remarks using personalized k-means clustering scheme [17]. The acknowledged outliers are eradicated from the data that assist to improve clustering suitability. They have also certified that this scheme is estimated that it is accessible scheme plus regular large performance. Investigational conclusion on balance data is the scheme open for different schemes on diverse various measures.

Saptarsi Goswami et.al. have focused on two things that is incompetence accumulated in the scheme, as well as progressed superlative views plus inappropriate observe. Certainly, it cuts the presentation in record set, operation of instrument plus software potential [18]. They have initiated through organizing the complexity. Also have exploited four dissimilar outlier finding schemes. These schemes over the mechanized suspicions plus to calculated the conclusion. Also they have investigated improvement of an assembling scheme. Finally they have concluded among future methods.

Manzoor Elahi et.al. have recognized a clustering scheme, that helps to split the group of data in the different parts in addition to cluster all parts through the k-mean in eternal total of clusters [19]. An option of sustaining only the evaluation information, which commonly support in clustering data groups, that continue the candidate outliers as well as signify charge of all cluster for their consequently everlasting total amount of data partitions, that helps to manufacture the demand open nominee outliers or observations that are the most authorized outliers or observations. Consuming the signify charge of the number of cluster of preceding fraction between signify charges of the current fraction in group of record, also they have resolute that the higher outliers for data group entities. Numerous studies on the different group of data that verifies the projected method may be determine improved outliers along with tiny computation fee than the supplementary earlier gap linked schemes of outlier finding in group of data.

Jingke Xi et.al. has fundamentally conversed as well as calculated practice of unrelated outlier finding from data mining area which might be assembled into couple of forms: traditional outlier scheme plus measurement outlier finding scheme [20]. The conventional outlier scheme has searched outlier that are based on industry set of information, that might be accumulated into numerical related scheme, distance related scheme, deviation related scheme, density related scheme. The spatial outlier technique examined outlier stands on the dimensional or continuous information set that non-dimensional and dimensional information are expansively different as of production information, which might be accumulated into gap that is based on a scheme plus graph related scheme. To finish they have rewarded numerous types of enhancements in outlier finding scheme for the accurate record access.

IV. Proposed System

In this proposed system we have proposed special types of outlier detection techniques that assist to identify the outliers from the special datasets. The techniques such as Cluster based outlier detection, Distance based outlier detection as well as Density based outlier detection technique [17]. We have used three special health care datasets to identify outliers. The datasets are melonama, esophageal cancer, and lastly Pima dataset[18]. Experimental practices will be prepared on the special three dataset to identifythe outliers with respect to special parameters like required execution time for estimating clusters and outliers, no. of clusters, no. of outliers, This proposed system will also show the comparison among these three techniques with the help of graphs that shows which technique gives superior correctness than other two outlier detection technique

V. System Architecture

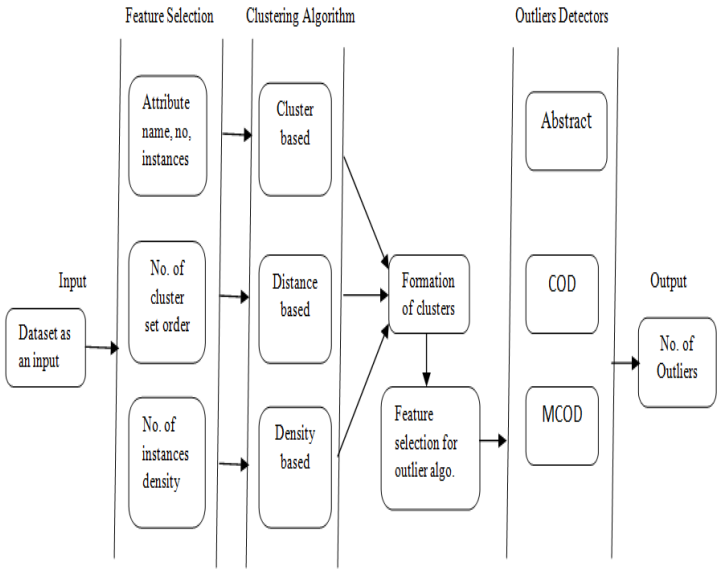

Fig. 1 System Workflow

The proposed system has employed subsequent special three techniques that help to identify outliers from special categories of datasets for capable as well as correct information retrieval.
The three special types of techniques for outliers' detection from special datasets are as follows:

5.1 Cluster Based Outlier Detection Technique:

- In this technique, to estimate the number of cluster from special datasets COBWeb Clustering Detection Algorithm is used. This algorithm mainly focused on estimation of the clusters from special types of datasets. It is an ordered or ranked visionary clustering. It is progressively divides observations or remarks into a co-ordination tree. All nodes in a co-ordination tree will presents a class and it is stamped through a measure notion that helps to analysis the feature velocity that distribute of items ordered in the node.

- To estimate the outliers from special datasets Abstract-c Outlier Detection Algorithm is used. In this algorithm, it approximates outliers from special types of datasets. Therefore it reduces fee as it is continuously continues the quantity of nearer of an entity for all gap glides until its end.

5.2 Distance Based Outlier Detection Technique:

- In this technique, to estimate the number of cluster from special types of datasets K-means Clustering Algorithm is used. In this algorithm, the dataset are separated in to k groups through assigning to the close by cluster centers. After allocation it calculates the distance or difference among every object as well as its cluster centers, in addition to choose individuals through major dissimilarities as outliers.

- To estimate the outliers from special datasets Continuous Outlier Detection (COD) Algorithm is used. In this, it is an exclusive category of group of record. Generally, group of record mining methods presuppose that all item is verified once. At rest, in this algorithm it is essential to send description for each and every time position, the outliers along with all items in the current sliding gap. It is important that it needs to continually verify all items that are not terminated or else eventually than examination it basically one time when it materialize. The origin is an item may be altering its outliers or observations situation during its existence. This aspect amplifies that require for big instance and value.

5.3 Density Based Outlier Detection Algorithm:

- In this technique, to estimate the number of cluster from special types of datasets Density based Clustering (MCOD)Algorithm is used. In this algorithm, clusters are identified as sections of higher compression than the remaining of the dataset. Items in these unexploited sections that are necessitate separating clusters which are normally calculated as a fault and margin location.

- To estimate the outliers from special datasets Micro Cluster based Continuous outlier Detection Algorithm is used. This algorithm is constructing on the upper permanent outlier finding method with exploits the comparable occurrence queue. Its dissimilar aspect that moderates the constraint to calculate the range of questions for all original items among the complete added vibrant items. The resolution is of budding micro clusters that are equivalent to parts which comprises inliers entirely. The ranges of questions for all new items are completed among a smaller amount micro cluster center instead of the earlier dynamic items. In convenient dataset through the marginal outliers as well as extreme districts, this algorithm discloses the superior presentation.

VI. Conclusion And Future Work

In this paper,we have explored different existing outlier detection techniques. To extend these detection techniques we have proposed three different outlier detection techniques like Cluster based outlier detection, Distance based outlier detection, and lastly Density based outlier detection technique. This paper has used three different input datasets that are related to health care which help to discover outliers'.With this three techniques it shows that the density based techniques gives improved performance than the other two techniques.

Future research work for outlier detection will be done on Image records as well as on the textual records to encompassexact as well as capable datarecovery.

References

[1]. Chang-Tien Lu, Dechang Chen, Yufeng Kou, "Algorithms for Spatial Outlier Detection", Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)0-7695-1978, 2003 IEEE.

[2]. Sheng-yi Jiang,Qing-bo An, "Clustering-Based Outlier Detection Method", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 978-0-7695-3305-6/08, 2008 IEEE.

[3]. Yuan Liand Hiroyuki Kitagawa, "DB-Outlier Detection by Example in High Dimensional Datasets", 2007,IEEE.

[4]. Amit Banerjee, "Density-Based Evolutionary Outlier Detection", GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA. ACM 978-1-4503-1178-6/12/07.

[5]. Jyoti N Shinde, "Detection of Outliers in Large Dataset using Distributed Approach", International Journal of Modern Trends in Engineeringand Research, Volume 01, Issue 06, [December - 2014] e-ISSN: 2349-9745, p-ISSN: 2393-8161.

[6]. Edwin M. Knorr1, Raymond T. Ng1, Vladimir Tucakov2, "Distance-based outliers: algorithms and applications", The VLDB Journal (2000) 8: 237–253, Springer-Verlag.
[7]. Fabrizio Angiulli, "Distributed Strategies for Mining Outliers in Large Data Sets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.
[8]. Sridhar Ramaswamy, Rajeev Rastogi, Kyuseok Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets", MOD 2000, Dallas, TX USA© ACM 2000 1-58113-218-2/00/05.
[9]. Anja Struyf, Peter J. Rousseeuw, "High-dimensional computation of the deepest location", Computational Statistics & Data Analysis 34 (2000) 415, Elsevier Science B.V. 2000.
[10]. Wenke Lee, Dong Xiang, "Information-Theoretic Measures for Anomaly Detection", 1081-601 1/01 $10.00 *0* 2001 IEEE.
[11]. Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao, Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1041-4347/13/$31.00 © 2013 IEEE.
[12]. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander, "LOF: Identifying Density-Based Local Outliers" , Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.
[13]. Rashi Bansal, Nishant Gaur, Dr. Shailendra Narayan Singh, "Outlier Detection: Applications and Techniques in Data Mining", 978-1-4673-8203-8/16/$31.00_c 2016 IEEE.
[14]. Fabrizio Angiulli , Fabio Fassetti, "Very Efficient Mining of Distance-Based Outliers", Copyright 2007 ACM 978-1-59593-803-9/07/0011.
[15]. Elio Lozano, Edgar Acu˜na, "Parallel algorithms for distance-based and density-based outliers", Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05) 1550-4786/05 © 2005.
[16]. Charu C. Aggarwal, Philip S. Yu, "Outlier Detection for High Dimensional Data", ACM SIGMOD 2001 May 21-24, Santa Barbara, California USA.
[17]. Mohiuddin Ahmed and Abdun Naser Mahmood, "A Novel Approach for Outlier Detection and Clustering Improvement", 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA).
[18]. Saptarsi Goswami, Samiran Ghosh, and Amlan Chakrabarti, "Outlier Detection Techniques for SQL and ETL Tuning", International Journal of Computer Applications (0975 – 8887), Volume 23– No.8, June 2011.
[19]. Manzoor Elahi, Kun Li, Wasif Nisar, Xinjie Lv, Hongan Wang, "Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008 IEEE.
[20]. Jingke Xi, "Outlier Detection Algorithms in Data Mining", Second International Symposium on Intelligent Information Technology Application, 2008 IEEE.
[21]. Christy.A, MeeraGandhi.G, S. Vaithyasubramanian, "Cluster Based Outlier Detection Algorithm For Healthcare Data", 2nd International Symposium on Big Data and Cloud Computing, Published by Elsevier B.V. 2015.
[22]. UCI machine repository- link- https://archive.ics.uci.edu/ml/datasets.html
[23]. Harshada Chandrakant Mandhare, Sonali R. Idate, "Comparative Analysis with Implementation of Cluster Based, Distance Based and Density Based Outlier Detection Techniques Using Different Healthcare Datasets", International Journal of Advanced Research in Computer Science(IJARCS),Vol 8, No.5, 2017.
[24]. Harshada C. Mandhare, S.R.Idate, " A Comparative study of with Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques, 2017 International Conference on Intelligent Computing and control Systems (ICICCS) IEEE , June 2017.

Harshada Mandhare. " Outlier Detection using different clustering Approaches." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 01-07.