ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Comparative Analysis with Implementation of Cluster Based, Distance Based and Density Based Outlier Detection Techniques Using Different Healthcare Datasets

Harshada C. Mandhare
M.tech, Student, Department of Information Technology,
Bharati Vidyapeeth University, College of Engineering,
Pune, India

Prof. S. R. Idate
Asso.Professor, Department of Information Technology,
Bharati Vidyapeeth University, College of Engineering,
Pune, India

Abstract: Outliers is view as an error data in information which is turned into important crisis that has been investigated in various areas of study plus functional fields. Several outlier detection methods have been implemented to assured functional fields, whereas several methods are supplementary basic. Various functional areas are also investigated in severe privacy like study on offense as well as terrorist behaviors. Through the improvement in information skills, the numeral of records, plus their measurement as well as difficulty, raise fast, that outcome in the need of computerized examination of huge quantity of various ordered data. For this intention, different data mining systems are utilized. The objective of these types of systems is to detect unseen dependencies from the records. Outlier detection in data mining is the detection of objects, remarks or observations that doesn't match to a predictable sample in a set of record. This detection technique is more beneficial in the several areas such as health trade, offense finding, fake operation, community protection and so on. In this paper we have studied different outlier detection algorithms such as Cluster based outlier detection, Distance based outlier detection plus Density based outlier detection. Result experimentation is done on different four dataset to identify the outliers and the comparative result shows that the cluster based methods are efficient for calculation of clusters and density-based outlier detection algorithm offers improved accuracy and faster execution for identification of outliers than other two outlier detection algorithm.

Keywords: Outliers; outlier detection; clusters; datasets; distance based.

I. INTRODUCTION

In the data mining different insignificant techniques is used to discover legitimate, original, and valuable as well as to end with clear samples. Nowadays, data mining is an essential implement which help to translate the records into information. This implementation essentially utilized in fraud discovery, selling in addition to technical finding. Data mining is essentially aid to removing the unseen remarkable samples from the huge quantity of datasets plus records. One of the major motivations of the data mining is to successfully as well as competently evaluate the group of the different remarks with their activities. In category to finish so, different cluster methods are enormously higher feature option. Clustering is the groups of remarks depending below single cluster are dissimilar in several mind since the further cluster. It is an unconfirmed learning method which is strictly aims at discovering the extreme plus spare areas in the record [1]. A variety of algorithms for data mining is a cascade record that doesn't robust in chief memory appropriate to requirement of assets where as this kind of huge record, the present data mining schemes are not that satisfactory plus prepared to contract amid them. A scheme that is implemented since cascade data for clustering is a suitable technique for managing vast quantity of modifiable record. Cascade data clustering is an important job in mining data cascades, the clustering can be measured the mainly vital unverified education difficulty; clustering is also identified as collecting associated items in the cluster. Via utilizing clustering technique we can able discover different outliers, so it has turn into one of the best data mining procedure therefore it is recognized as outlier mining [2]. Outlier detection techniques mainly classified which is mainly depending on the accessibility of guidance record. These types of techniques additionally reduce into different three groups: supervised technique, semi-supervised technique, as well as unsupervised technique. In general, supervised techniques have a big accuracy speed. Still, they need pre-tagged data that is labeled as a usual or outlier. In calculation, supervised techniques are appropriate for records whose character doesn't modify during instance. Semi-supervised techniques are lying on the accessibility of guidance record set of normal outlier remarks. Semi-supervised techniques could imperfectly classify a normal outlier remark which reduces external the qualified margin, like an outlier remark. The restriction of supervised as well as semi-supervised techniques which is the education record set should characterize every potential set. In the unsupervised techniques, it practices data with no several previous understanding. The benefit of unsupervised techniques is nothing but no tagged data is necessary. Still, unsupervised outlier discovery techniques typically have greatly superior copied distress speeds than the copied distress speeds of supervised as well as semi-supervised techniques [3].

The previous is a circumstance where groups of record set have information regarding the category of items which is regular or irregular, plus the last doesn't have category data. An original technique for outlier discovery is via spotlighting on the unsupervised case study. While outlier discovery could be concerned to different areas like as interruption discovery, deception discovery, error discovery, physical condition checking organizations, as well as discovery of ecology turbulence, numerous discovery techniques has been projected [4]. The benefits of the

limited outlier issue that advance above additional outlier discovery methods are: It identifies outliers among value to thickness of their adjoining records; that is not to the worldwide form plus it is capable to discover outliers despite the record sharing of regular activities, as it doesn't build some suppositions regarding the sharing of records. Methodology which helps to decreases the local outlier factor algorithm's calculation instance. To decrease the calculation instance in the local outlier factor computation, kd-tree indexing plus an estimated adjacent national algorithm are used [5]. Input characteristic division technique is used to permit to discover every outlier on the complete characteristic group, once that it turn into probable to explore during every remote quality divisions for these positions. As a result, the information characteristic subspace could be recognized via computing the remote divider connection amid every remote characteristic division plus the complete characteristic group. In this outcomes illustrate that this method can be resourcefully useful on huge dimensional group of data for outlier discovery [6]. Radiance weight ordering method is motorized via region perceptive muddling, that re-ranks the record points via their possibility of individual an outlier. Academic disagreements that validate the underlying principle for the advance plus afterward it perform a widespread experiential lesson that importance the efficiency of method within reach of over existing results [7]. Outlier aspects of item to the case of cluster plus set onward a clustering stands outlier discovering technique. The cluster comes via clustering procedure as a division that categorizes as it is regular otherwise outlier observation [8]. An algorithm called density stands trajectory outlier discovery that utilizes the benefit of trajectory outlier discovery. As this discovery not able to identify limited outliers successfully, then scheme density is utilized [9]. In the density stands outlier discovery technique, the local outlier factor discovers outlier through evaluating the items from its local density to its nearer local density. The regular items local density is estimated to its nearest, as the outlier items local density is considerably minor than it's nearer. It calculates the items local density via calculating the remoteness among items, and also discovers the outlier [10].

In this paper we have implemented novel different outlier detection algorithms namely, Cluster based outlier detection, and Distance based outlier detection as well as Density based outlier detection. For execution of these algorithms we have used four different dataset to discover the outliers and also show comparison between these three algorithms that shows the cluster-based outlier detection algorithm gives enhanced accuracy than other two outlier detection algorithm.

This paper gives special characteristics:

- It helps to evaluates different types of outlier detection algorithms with the help different parameters such as no. of clusters, no. of outliers as well as execution time for calculating outliers and clusters.
- It comparatively analyzes result achieved via three different techniques that shows which technique gives better efficiency and accuracy.

The paper is prepared are as follows: In section 3 we have explored the earlier different big data handling techniques which helps to discover outliers. In section 4 we have Implemented novel three different algorithms which help to discover the outliers to retrieve accurate data. In section 5 we have define the system architecture. In section 6 shown experiment result of these three algorithms plus comparative performance of three different algorithms with the help of graphs and tables. Illustrate a conclusion and future scope in data mining system in section 7.

II. BACKGROUND AND MOTIVATION

In several data analysis jobs a huge amount of characteristics are being traced or exampled. One of the primary tasks in the direction of gaining a consistent examination is the discovery of expending observations or remarks. Even if outliers are frequently measured as a fault, they might bring vital data. Discovered outliers are applicant for abnormal record that might or else unfavorably direct to copy non-measurement, subjective limited evaluation plus wrong outcomes. Here an outlier is a nothing but observation or remark point which is isolated since additional observations or remarks. An outlier might be owed to inconsistency in the dimension or it can specify investigational fault; latter it disqualified from the different dataset. Outlier discovery is suitable in a different of areas, like interruption discovery, deception discovery, error discovery, physical condition checking, incident discovery in sensor systems, as well as discovering Eco turbulence. It is frequently utilized in pre-executing to eliminate irregular record from the different dataset. As an outcome, they able bias as several study performed on the different dataset. It is consequently significant to discover plus sufficiently covenant with outliers or remarks.

An outlier in a different protection background, a deception discovery system, an image investigation system or an interruption checking system should be discovered instantly in a real time plus an appropriate distress echo to aware the system supervisor to the difficulty. Formerly the circumstances has been griped, this irregular interpretation might be accumulated individually for evaluation among several original deception problems but would possibly not be accumulated among the major system record as these different methods lean to representation regularity plus utilize this to discover outliers. Therefore to avoid different fraud problems outliers' detection from the different data set is extensively important to access accurate and important data.

III. RELATED WORK

Mohiuddin Ahmed and Abdun Naser Mahmood have explored a new unsupervised technique to identify outliers or remarks with the help of customized k-means clustering algorithm [11]. The identified outliers are eliminated from the record set which helps to progress clustering correctness. They have validated this approach via evaluating beside offered methods as well as standard presentation. Their experimental outcomes on scale record sets illustrate that the method complete offered techniques on different numerous procedures.

Ana arribas-gil and Juan room have proposed an innovative technique to imagine and identify outline outliers in trials of arcs [12]. In practical data investigation, they have viewed arcs that are described above a specified valid distance plus outline outliers might be classified as those arcs that show a

dissimilar outline from the remaining of the trial. While significance outliers, that is, arcs that stretch out slight the different mainstream of the record, that are effortless to recognize, outline outliers are frequently covered amid the remaining of the arcs moreover so complex to discover. In this article, they have exploited the association among two procedures of intensity for practical record that assist to imagine arcs in the form of outline and also to develop an algorithm for outline outlier discovery. Also they have shown the use of visualization software tool, the outlier gram, during numerous cases and explore the algorithm performance on the recreation learning. Lastly, they have applied method to evaluate cluster superiority in a valid group of array data.

Saptarsi Goswami et.al. In this paper, they have not only point out ineffectiveness assembled in the system, but also they have evolved best observes and unsuitable practice. Undoubtedly that decrease performance in data stream plus deployment of hardware as well as software capability. In this paper they have started with preparing the difficulty. They have used four different outlier discovery methods. These methods above corpora of manufacturing doubts in addition to evaluated the outcomes. They have also explored advantage of an assembly method [13]. Lastly they have concluded with upcoming courses of accomplishment.

Bo Liu et. al. This paper has presented an original outlier discovery method that addresses record through damaged tags as well as integrates restricted irregular paradigms into knowledge [14]. To contract with record along with damaged tags, they have introduced odds rates for every input record that indicate the grade of relationship of an model near the regular as well as irregular groups correspondingly. Proposed method performs in two stages. In first stage, they have produced a simulated education record set via calculating probability rates of every model stands on its restricted activities. They have shown kernel k-means clustering technique and kernel local outlier factor stands on technique to calculate the probability rates. In second stage, they have integrated the produced probability rates plus restricted irregular patterns into education structure to construct an added precise classifier for worldwide outlier discovery. Via combining restricted as well as worldwide outlier discovery, proposed technique unequivocally holding record through damaged tags plus improves the presentation of outlier discovery. Widespread trials on real life record groups have confirmed that the proposed methods can accomplish a recovered exchange amid discovery speed plus artificial distress speed as balanced to outlier discovery methods.

Manzoor Elahi et.al. In this paper they have established a clustering related method, which separate the record set in portions plus cluster every portion via k-mean in permanent amount of clusters [15]. As an alternative of maintaining simply the review record, that frequently aid in case of clustering record set, that remain the applicant outliers plus mean rate of each cluster for the subsequently permanent amount of record portions, to make definite that the discovered applicant outliers are the valid outliers. With utilizing the mean rate of the clusters of prior portion among mean rates of the present portion of record set, they have decided the superior outliers for record set items. Several researches on dissimilar record set verify that the proposed technique can discover enhanced outliers among little calculation charge than the additional exiting space related methods of outlier discovery in record set.

Jingke Xi et.al. This paper has essentially discussed and evaluated process of dissimilar outlier discovery from data mining perception that can be grouped into two types: classic outlier technique as well as spatial outlier technique [16]. The classic outlier technique has explored outlier stands on business record group, which can be assembled into statistical stands technique, distance stands technique, deviation stands technique, density stands technique. The spatial outlier technique examined outlier stands on the spatial record group that non spatial as well as spatial record are extensively dissimilar from business records, that can be assembled into space stands technique as well as graph stands technique. At last, this paper has concluded some progresses in outlier discovery newly.

IV. PROPOSED SYSTEM

In this section we have studied different types of outlier detection algorithms which help to detect the outliers from the different datasets. Algorithms are Cluster based outlier detection algorithm, Distance based outlier detection algorithm plus Density based outlier detection algorithm [17]. Four different input datasets are used for these algorithms to detect outliers are melonama dataset, esophageal cancer dataset, Pima dataset as well as Diabetes dataset [18]. Experimental practices is done on different four dataset to detect the outliers with the help of different parameters such as no. of clusters, no. of outliers, execution time for calculating clusters and outliers and the comparison between these algorithm are shown to define which algorithm provides enhanced accuracy than other two outlier detection algorithm.

V. SYSTEM ARCHITECTURE

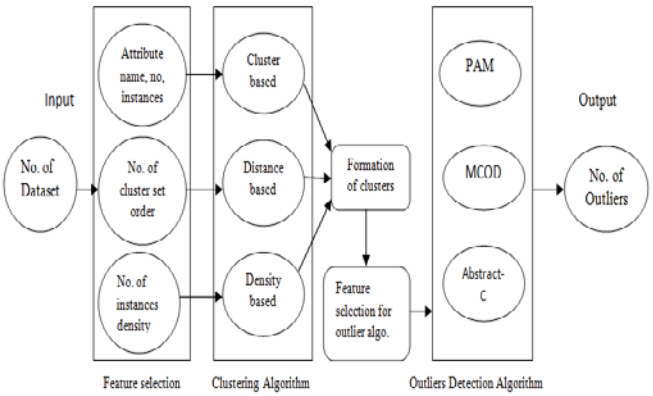

Fig. 1 System Workflow

The proposed system has used following different three algorithms to detect outliers from different types of datasets for efficient and accurate data retrieval are as follow:

A. Cluster Based Outlier Detection Algorithm:

1) COBWeb Clustering Detection Algorithm:

This algorithm is used to estimate the clusters from different types of datasets. This algorithm is a hierarchical conceptual clustering. It is increasingly

categorizes remarks into a classification tree. Every node in a classification tree shows a class as well as is tagged by a quantity conception that reviews the characteristic rate sharing of objects organized in the node.

2) Partition Around Medoids Outlier Detection Algorithm:

In partition around medoids algorithm, it helps to estimate outliers from different types of datasets. This algorithm is used to discover an order of items which is known as a medoids that is located in the midway of clusters. Items are uncertainly characterized as a medoids that are located into group of preferred items.

B. *Distance based outlier detection algorithm*

1) K-means Clustering Algorithm:

This algorithm is used to calculate the cluster from different datasets. In the k-means algorithm the dataset are divided in to k factions via conveying them to the nearby cluster hubs. After allocation it calculates the distance or difference among every object as well as its cluster hub, and selects those with biggest differences as outliers.

2) Micro Cluster based Continuous outlier Detection Algorithm:

This algorithm is used to detect the outliers from different datasets. It is built on top of continuous outlier detection algorithm and utilizes the similar event queue. Its distinct attribute which mitigates the requirement to estimate variety queries for every new object with the entire additional dynamic objects. The solution is of developing micro clusters which match to areas including inliers completely. Then the variety of queries for every original object is executed with less micro cluster hub alternatively of the preceding active objects. In practical dataset with the minority outliers plus intense areas, this algorithm reveals the improved performance.

C. *Density based outlier detection algorithms*

1) Density based Clustering Algorithm:

This algorithm is used to calculate the clusters from the datasets. In this algorithm, clusters are defined as regions of superior compactness than the rest of the dataset. Objects in these unused regions which are need to divide clusters –that are typically measured to be error as well as border position.

2) Abstract-c Outlier Detection Algorithm:

In this algorithm, it helps to estimate outliers from different types of datasets. It decreases cost as it constantly maintains the numeral of neighbors of an object for every window slides awaiting its finish.

VI. EXPERIMENTAL RESULT

In this section, we have done experimental practice on four different health care dataset to detect the numbers of outliers. The primary dataset is melanoma, the secondly dataset is esophageal cancer dataset and third dataset is Pima and the last dataset is Diabetes. Performance of three different outlier detection algorithms are shown with the help of graphs are as follows:

A. *DATASET 1 (MELONOMA)*

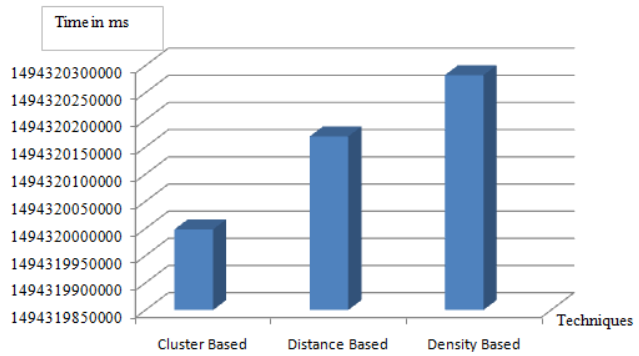

Fig.1 Parameter- Execution time for calculating clusters

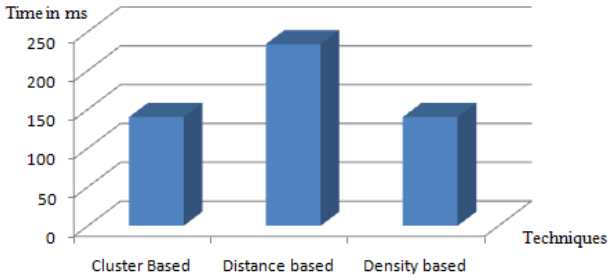

Fig.2 Parameter- Execution time for calculating outliers

Fig. 3 Parameter - No. of clusters

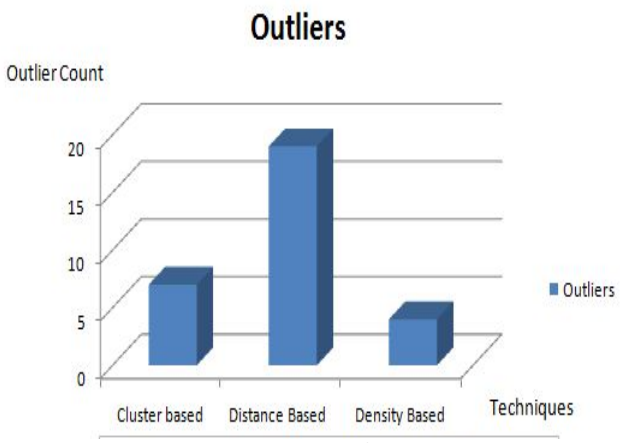

Fig. 4 Parameter- No. of outliers

B. *DATASET 2 (ESOPHAGEAL CANCER)*

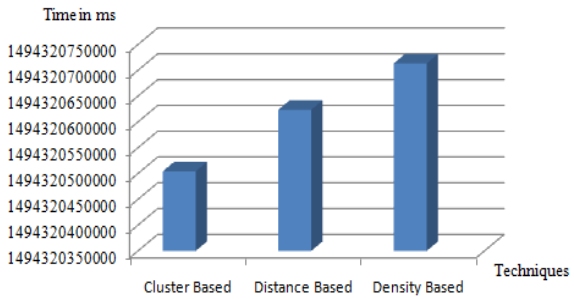

Fig. 5 Parameter- Execution time for calculating clusters

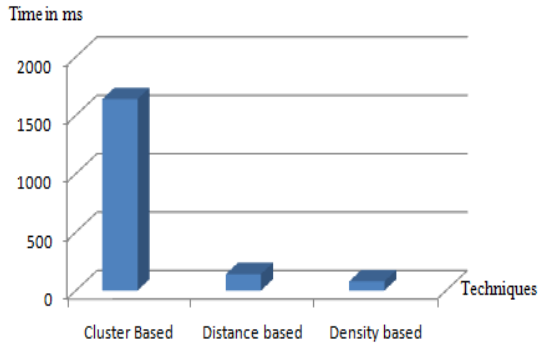

Fig. 6 Parameter- Execution time for calculating outliers

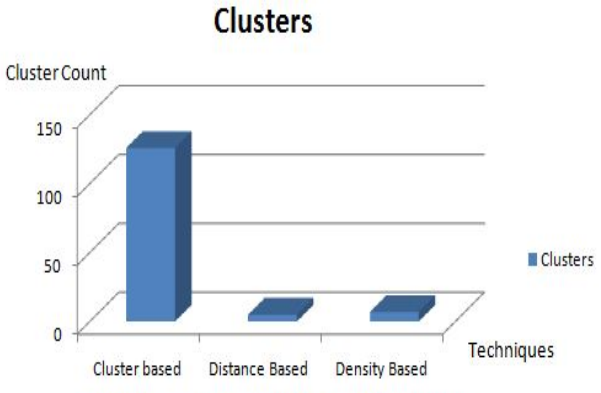

Fig. 7 Parameter- No. of clusters

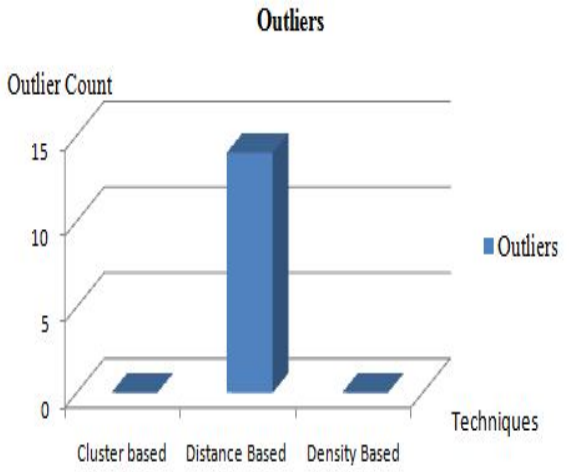

Fig. 8 Parameter- No. of outliers

C. *DATASET 3 (PIMA)*

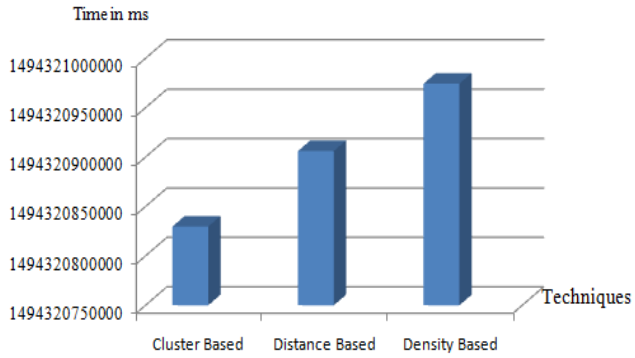

Fig. 9 Parameter- Execution time for calculating clusters

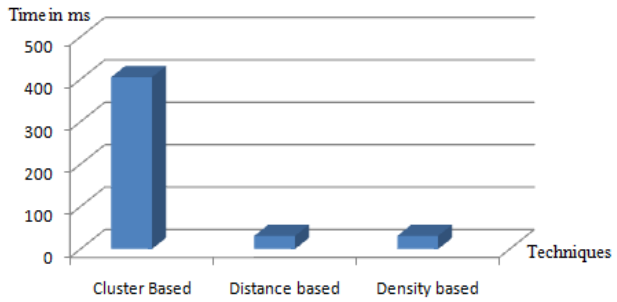

Fig. 10 Parameter- Execution time for calculating outliers

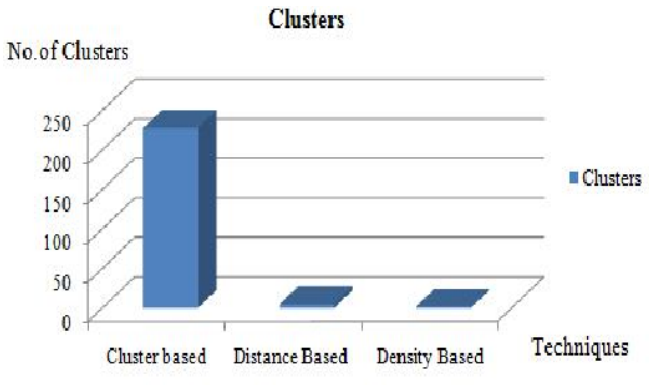

Fig. 11 Parameter- No. of clusters

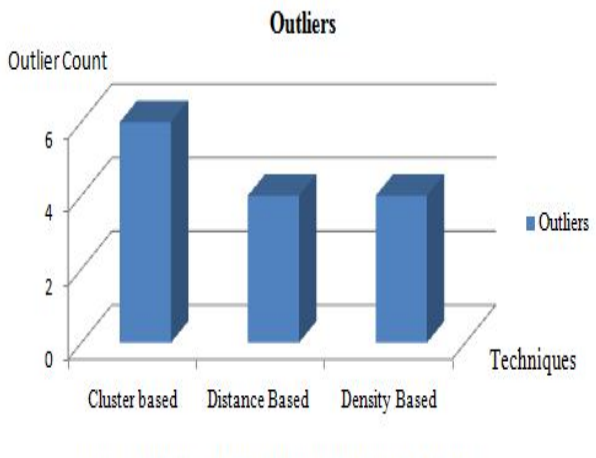

Fig. 12 Parameter- No. of outliers

D. DATASET 4 (DIABETES)

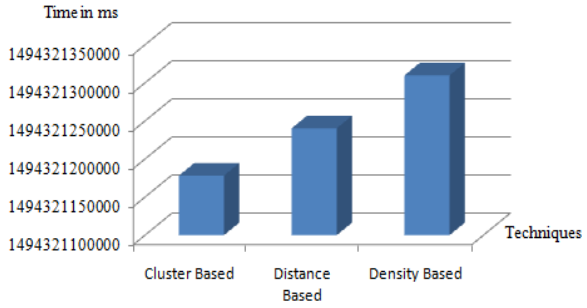

Fig. 13 Parameter- Execution time for calculating clusters

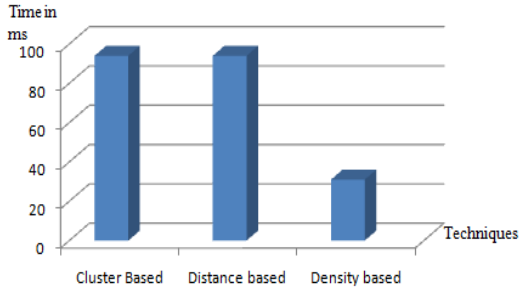

Fig. 14 Parameter- Execution time for calculating outliers

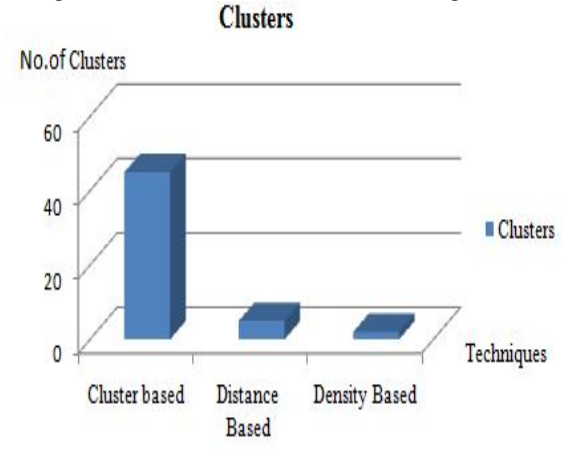

Fig. 15 No. of clusters

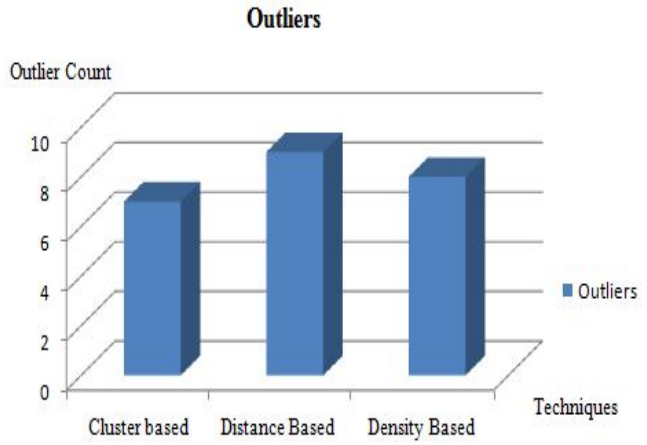

Fig. 16 Parameter- Execution time for calculating outliers

The parameters required for the experimentation on the dataset is illustrated on four different Tables.

Table.1 Dataset 1(Melanoma)

Dataset 1-Melonema				
	No. of Clusters	**Execution Time to Calculate Clusters**	**Execution Time to Calculate Outliers**	**No. of Outliers**
Cluster Based	**241**	**1494319998428 ms**	**140 ms**	**7**
Distance Based	**5**	**1494320169465 ms**	**234 ms**	**19**
Density Based	**8**	**1494320282248 ms**	**140 ms**	**4**

Table.2 Dataset 2 (Esophageal Cancer)

Dataset 2- Esophageal cancer				
	No. of Clusters	**Execution Time to Calculate Clusters**	**Execution Time to Calculate Outliers**	**No. of Outliers**
Cluster Based	**126**	**149432050421 ms**	**1639 ms**	**0**
Distance Based	**5**	**1494320623399 ms**	**140 ms**	**14**
Density Based	**7**	**1494320713390 ms**	**78 ms**	**0**

Table.3 Dataset 3 (Pima)

Dataset 3- Pima				
	No. of Clusters	**Execution Time to Calculate Clusters**	**Execution Time to Calculate Outliers**	**No. of Outliers**
Cluster Based	**228**	**1494320829672 ms**	**405 ms**	**06**
Distance Based	**5**	**1494320906123 ms**	**31 ms**	**04**
Density Based	**2**	**1494320974304 ms**	**31 ms**	**04**

Table.4 Dataset 3 (Diabetes)

Dataset 4 - Diabetes				
	No. of Clusters	**Execution Time to Calculate Clusters**	**Execution Time to Calculate Outliers**	**No. of Outliers**
Cluster Based	**45**	**1494321178122 ms**	**94 ms**	**07**
Distance Based	**5**	**1494321240312 ms**	**94 ms**	**09**
Density Based	**2**	**1494321309210 ms**	**31 ms**	**08**

VII. CONCLUSION AND FUTURE WORK

In this paper, we have implemented three different outlier detection algorithms such as Cluster based outlier detection algorithm, Distance based outlier detection algorithm and Density based outlier detection algorithm. Four different input datasets are used for these algorithms to detect outliers such as melanoma dataset, esophageal cancer dataset, Pima dataset and Diabetes dataset. Experimentation done on the different four dataset to detect the clusters and outliers in addition to it also shows the comparison of these three algorithm with the help of graphs that demonstrate cluster based outlier detection algorithm gives better performance to calculate the number of clusters as well as density based outlier detection algorithm gives better performance to calculate the number of outliers.

A one of motivating future research work can be done on Image datasets as well as on the text file datasets to have accurate plus efficient information retrieval.

VIII. ACKNOWLEDGE

This work is supported by the Prof. S. R. Idate Associate Professor, Department of Information Technology, Bharati Vidyapeeth University, College of Engineering, Pune, India.

IX. REFERENCES

[1] Kamal Malik, H.Sadawarti, Member IEEE, Kalra G.S., Member IEEE," Comparative Analysis of Outlier Detection Techniques", International Journal of Computer Applications (0975 – 8887) Volume 97– No.8, July 2014.

[2] Dr. S.Vijayarani, Ms. P. Jothi, "Comparative Analysis of Clustering Algorithms for Outlier Detection in Data Streams", International journal of engineering sciences & research technology, issn: 2277-9655, 2013.

[3] Armin Daneshpazhouh, Ashkan Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples" , 0167-8655, 2014 Elsevier B.V.

[4] Jihyun Ha, Seulgi Seok, Jong-Seok Lee, "Robust outlier detection using the instability factor" , Knowledge-Based Systems 63 (2014) 15–23, _ 2014 Elsevier B.V.

[5] Seung Kim, Nam Wook Cho, Bokyoung Kang, Suk-Ho Kang, "Fast outlier detection for very large log data", Expert Systems with Applications 38 (2011) 9587–9596, 2011 Elsevier Ltd.

[6] Peng Yang, Qingsheng Zhu, "Finding key attribute subset in dataset for outlier detection", Knowledge-Based Systems 24 (2011) 269–274, 2010 Elsevier B.V.

[7] Ye Wang, Srinivasan Parthasarathy, Shirish Tatikonda, "Locality Sensitive Outlier Detection: A Ranking Driven Approach", ICDE Conference 2011, IEEE.

[8] Sheng-yi Jiang,Qing-bo An, "Clustering-Based Outlier Detection Method", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008 IEEE DOI 10.1109/FSKD.

[9] Zhipeng Liu, Dechang Pi, and Jinfeng Jiang, "Density-based trajectory outlier detection algorithm", Journal of Systems Engineering and Electronics Vol. 24, No. 2, April 2013, pp.335–340.

[10] Haowen Guan, Qingzhong Li, "SLOF: Identify Density-based Local Outliers in Big Data", 2015 12th Web Information System and Application Conference , IEEE DOI 10.1109/WISA.

[11] 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).

[12] Ana arribas-gil and Juan romo, "Shape outlier detection and visualization for functional data: the outliergram", Biostatistics Advance Access published March 11, 2014.

[13] 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.

[14] 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.

[15] 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.

[16] Jingke Xi, "Outlier Detection Algorithms in Data Mining", Second International Symposium on Intelligent Information Technology Application, 2008 IEEE.

[17]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.

[18]UCI machine repository- link- https://archive.ics.uci.edu/ml/datasets.html