Middle-East Journal of Scientific Research
24 (Special Issue on Innovations in Information, Embedded and Communication Systems): 216-223, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.IIECS.23163

Outlier Detection for Uncertain Data Stream Having Existential Probability

K. Masila devi and R. Karthikeyan

Department of CSE, PSNA College of Engineering and Technology, Dindigul, India

Abstract: In recent days real time data set consist of many noisy and unwanted data. So it is quite difficult for the further processing of the data. Detecting outlier is considered to be the primary step in many data-mining applications. An outlier is also known as anomaly which is significantly deviated from the original data. It is essential to remove that outlier for the better knowledge discovery. Data may also be continuous, such data are known as data streams. For those data stream, the outlier detection is very difficult and is detected continuously. Due to the proliferation of uncertain data in the data streams, the anomaly is detected at a fly. Each data is considered to be a tuple. A tuple may or may not present in the database and the presence and absence of one tuple may affect the probability of the presence or absence of another tuple in the database. This is called as an existential probability of uncertain data stream. A novel distance based outlier detection used for detecting outliers in objects with assigned existential probability. As the data are continuous, sliding window is used to consider the most recent data. The probability of the presence and absence are calculated using probabilistic functions. This technique outperforms the existing approach and helps in improving the efficiency of the system. The performance will be examined using the real world data to verify the theoretical study of the algorithm.

Key words: Clusters · Outlier detection · Probability · Tuple

INTRODUCTION

Outlier Detection: Outlier detection is to find patterns that do not match to expected behavior. Outliers are patterns in data that do not belong to the normal behavior of the data. Also an outlier in a set of data that appears to be deviated with the remainder of the data in the same set. Outliers arise due to various reasons like human mistake, instrument blunder, regular deviations in populations, fraudulent behavior and also due to the changes in behavior of systems or faults that occur in the systems, etc. An outlier can be either a single data point or a cluster of data. The approaches used in outlier detection are,

- To detect the outliers with having no prior knowledge about the data that is present. This learning approach is similar to unsupervised learning.
- To model both normal and abnormal behavior of the data an approach that is similar to supervised learning is used and it requires pre-labeled data that are tagged as normal or abnormal data which is similar to semi-supervised learning

The several issues in the prediction are dimensions of the characters, data redundancy, missing of data, skews in the data, invalid data and so forth. In uncertain data management, data records are usually represented in terms of probability distributions rather than any deterministic values. Here detecting rare patterns using mining techniques are used to understand and detect the patterns of change based on the probabilistic values of the data.

Causes: Outliers exist in almost every real data set. Some of the causes for outliers are

Malicious Activity: Such as insurance or credit card fraud, a cyber intrusion, a terrorist attack

Instrumentation Error: Such as defects in parts of machines or wear and tear

Change in the Nature: Such as a environmental change, a new buying pattern among consumers, transformation in genes

Corresponding Author: K. Masila devi, Department of CSE, PSNA College of Engineering and Technology, Dindigul, India

Human Error: Such as an automobile accident or reporting the data error

Modelling Techniques: The commonly used approaches in recent times are, Model-based Approaches: In model based approach, a specific model is applied to the data points. And identify if the data is fit to that model or not. If any data that do not fit into that model is known as outliers. These outliers are removed to enhance the accuracy of the data. Some methodologies in this approach are,

- Probabilistic method
- Depth based approaches.
- Deviation-based approaches.

Proximity-based Approaches: The data in data space are analyzed to find the spatial proximity. If the proximity of an object is considered to deviate from the proximity of other objects then it is an outlier. Those outliers are detected and removed. The methodologies are,

- Distance-based approaches.
- Density-based approaches.

Angle-Based Approaches: Analyze the pair wise angles between a given point and all other points in the dataset. Outliers are points that fluctuate high. Those data points are eliminated from the data space.

Applications: Outlier detection has enormous applications, including detection of credit card fraud, to detect criminal behavior in E-commerce, video surveillance, pharmaceutical research examination, prediction of climatic changes and discovery of unexpected astronomical objects.

This paper is organized as follows, In Section II, the related survey about the paper is briefly discussed. In section III, the proposed system is explained with the detailed architecture. In the section IV, the experimental results are verified. Section V ends up with conclusion.

Related Work: The research on many outlier detection in data streams and uncertain data are studied here. Initially finding outliers in massive information set may be a tough task. One technique that's been applied to discover outliers in massive information set is, a way for selecting and deleting distance based mostly outliers in terribly massive information sets introduces parallel computation therefore like spare extra time and having brilliant execution. Initially, a graph is created supported the information set. Then weights are assigned to every of the data's within the graph. Algorithmic rule income by assignment weights to the information in every row. As per the load a threshold price is ready. If there's an opportunity that information exceeds the edge price, then that row is taken into account as nothing but an outlier and it's off from the information set. during this approach outliers from all the information sets are obtained. Then mechanically weights are assigned. And also the same procedure is going to be recurrent. By deleting the outliers, it will increase the memory for storing additional information [1].

Outlier Detection in information stream, Angle based framework is utilized to find the exceptions in assorted strategies which is more successful. List based calculation and Random hyper plane projection calculation are the two calculations utilized as the piece of Angle based strategy to discover the anomalies. List based calculation is set to discover indexing esteem between the articles. Irregular hyper plane strategy is utilized to locate the careful worth in the dataset arbitrarily. Sliding window technique is actualized to distribute the memory space for the dataset [2].

At that point for identifying exceptions in extensive database another calculation was presented. The strategy reports all anomalies by examining the dataset at generally twice. So in this paper another calculation, SNIF (scan with organized flushing) was presented. Bear to hold more questions in memory amid the principal dataset examine itself, which permits allows a noteworthy segment of R as non anomalies direct after the sweep. Subsequently, the remaining items that require further confirmation might fit in memory, so that another sweep of R suffices to decide the accurate anomalies. Taking into account this thought, SNIF conveys a novel organized flushing strategy to minimize the shot of performing the third output of R. In particular, the method relates each article with a "need" and, at whatever point the memory turns out to be full, flushes the items with the least needs [3].

BIRCH (Balanced Iterative Reducing and Clustering utilizing Hierarchies) and demonstrates that it is especially suitable for huge databases. This framework finds a decent quality bunches in a solitary output of information. Adjusted Iterative Reducing and Clustering utilizing Hierarchies is presented for huge databases. However, BRICH can't expand the limit powerfully. Furthermore, impractical to progressively modify exception criteria [4].

An iterative arbitrary testing methodology is produced that looks at separated of the whole information space. In the event that an article does not have a place with the analyzed space, it is viewed as an exception in the relating emphasis. As such,if an article is in that space,it is seen an inlier.By rehashing this system of space examination,the proposed technique figures the inlierness score, which is known as the Observability factor,for every item in a dataset. Since the OF shows the inlierness score, an article with a low OF quality is a promising anomaly competitor ideally. This strategy is identified with separation based methodologies, since by implication it uses the separations between objects [5].

Various exception recognition methods require client characterized parameters which require learning for the compelling operation over information streams. Dynamic way of information streams, the parameter qualities are difficult to anticipate and the parameter that is picked may not be suitable all through the lifetime of an information stream.

Since the climate information will be indeterminate it is imperative to recognize anomalies in such information stream. Exception discovery on unverifiable information streams which can empower the simultaneous execution of various inquiries at the same time. Despite the fact that it diminishes the required stockpiling overhead, more productive, Offer huge adaptability as to the info parameters [6].

Steady unverifiable exception location which can quickly choose the method for the uncertain segments by pruning to upgrade the viability. Pruning is to lessen the distinguishing proof expense. It is normal exception likelihood. Anomaly recognition on ceaseless information streams which can manage the parameter variable inquiries at the same time. Be that as it may, Only the progressions of k in (R, k,lambda) is contemplated this paper [7]. Various techniques rely on upon the fundamental qualities of specific information to recognize the outliers. So they can't be viably associated with differing sorts of information in other application spaces. They should be tuned and changed to conform to the new space. In this paper a diagram based technique for the revelation of relevant anomaly in consecutive information is proposed. The calculation offers a more elevated amount of flexibility and requires less measure of information about the way of the broke down information contrasted with past philosophies. The precision and productivity very depends on the info parameters and limits of the information stream. In this way Cannot naturally assess the qualities for various datasets [8]. Initially, a cell-based methodology of separation construct anomaly location in light of questionable articles taking after by the Gaussian dissemination is proposed. Second, an estimated cell-based methodology of anomaly recognition utilizing the limited Gaussian dispersion is proposed to expand the productivity of exception location. Guess in the Gaussian circulation beats the limited Gaussian dispersion empowering more powerful pruning. What's more, in this manner the level of vulnerability is limited [9]. Mining Outlier in Data Streams Using Massive Online Analysis Framework requires just very less time to execute. Be that as it may, just the conventional techniques are confirmed utilizing this framework [10].

Proposed System: In the proposed system the weather data set is considered. Weather data may consist of numerous noise and unclean data which are due instrumental error or any human error. So the data set is pre-processed initially. This phase is done to remove the data that are deviated from the rest of the data and it can be identified by the user. But some data are not noise, but greatly deviated from the rest of the data is called outliers.

Since the data are uncertain and continuous in nature normal outlier detection algorithm is not suitable for detecting anomalies. Moreover, the data are existential the outliers are detected using the probability of the data.

The sliding window concept, which tracks the most recent data in the data stream is used to detect the outliers. The data within the window is considered to be the active data. Then the data expires as soon as they leave the window. This window is triggered periodically to detect outlier in the data stream.

Since the data with the existential probability is considered, the data may influence the other data. So probability function is used to determine the presence or absence of the object. The data may influence the other data at the time of arrival and also at the time of departure. So the outliers are calculated at both the cases.

Now the clusters are formed. The outlier may not only be the single data point that is abnormal form the rest of the data but also can be the cluster of data. So the outliers are detected from the clusters formed based on the probability.

System Architecture: In the proposed system the uncertain data is given as an input and then it is passed to the various phases like pre-processing, clustering and outlier detection to find the outliers from the uncertain data stream. And finally the clusters and outliers are visually represented as illustrated in Fig. 1.

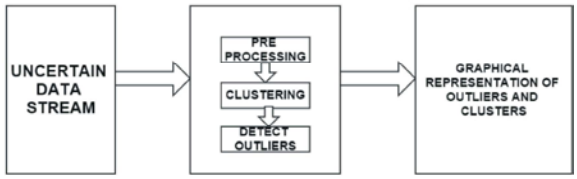

Fig. 1: Outlier Detection Architecture

The outlier from the input data stream is removed and the clusters are extracted efficiently as shown in the Fig. 2.

The uncertain weather data stream which has the existential probability is given as the input to the system. Here the data stream may have some noise or any missing values, which is to be removed initially. This technique is known as pre-processing. It is considered to be the foremost step in any data mining techniques as the noisy data may lead to wrong knowledge gain. The pre-processing steps are carried out like data cleaning, data reduction, data transformation, etc. According to the nature of the data the pre-processing is being carried out. Then the pre-processed data is further utilized for the clustering. Here the outliers are detected using the cluster formation. Since the data posses existential probability and are streaming in nature, the sliding window technique is used to capture the recent data and clustering is performed. The data are clustered based on the distance. So the Euclidean distance is calculated for the data within the window. And the window is moved periodically to calculate the probability for the data in the entire data set. The clusters are formed until the data becomes an outlier. So the same technique is repeated for the entire dataset. And the various clusters are obtained.

Having the clusters it is important to discover the anomalies from that. Here the outliers can be either a data point or any cluster. So the outliers are computed with the threshold value. The data which is less than the threshold forms the clusters. And any data or cluster that does not fit into the criterion is detected as an outlier.

The clusters and outliers are visually represented to examine the nature of the cluster. Now the data with all the outliers removed are obtained. This meteorological data can be used for further analysis like weather prediction, any astronomical computations.

Pre-Processing: Pre-processing is used to remove the noisy and unwanted data from the data stream. This is illustrated in the Fig. 3.

The initial and the foremost step in data pre-processing is that the data is analyzed to check for any irrelevant or noisy data or any missing values. After the proper analysis, the cleaning process is done.

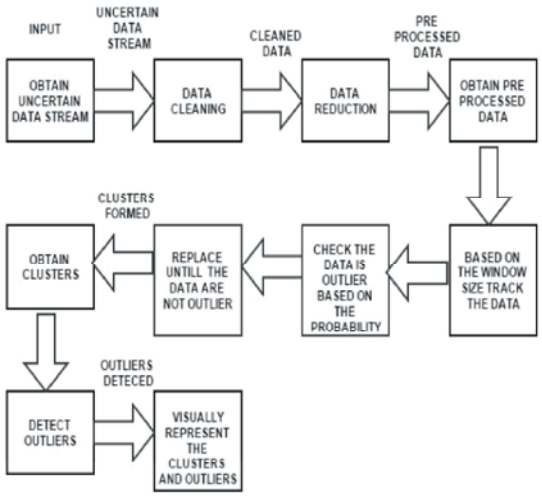

Fig. 2: Work Flow for Detecting Outliers

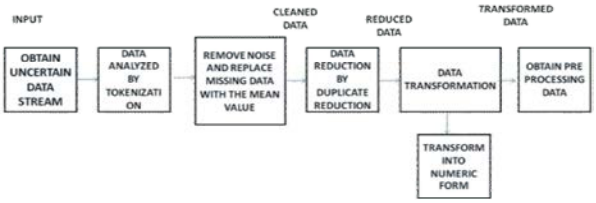

Fig. 3: Pre-Processing

There are many reasons for noisy data (having incorrect attribute values) in the data set. The instruments used for collecting data may be faulty. The data entry error that are caused due to human entry or computers. Data transmission errors can also occur. There may be limitations in technology, such as limited buffer size for transfer and consumption of synchronized data. The noise data can be eliminated by Binning methods smooth a sorted data value by consulting its neighborhood that is, the values that are around it or Data can also be smoothed by fitting the data to some function, like regression.

Data cleaning work to "clean" the data by filling in missing values as shown in the Fig. 4, smoothing noisy data, identify and remove the outliers and reduce inconsistencies. The missing values can be replaced in any of the methods like ignoring the tuple, fill in the missing value manually, replace with the mean value of the column. The attributes will be filled and no empty cell is left.

```
if (chrs >= 32 && chrs <= 47) {
append 0;
} else {
value;
}
```

The data reduction is carried out on the weather data set if they have any irrelevant, weak or redundant attributes or dimensions may be detected and removed using attribute subset selection.

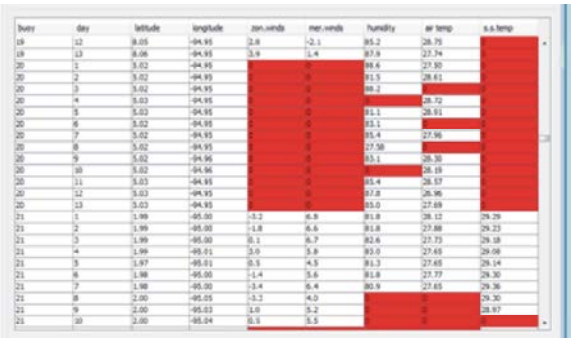

Fig. 4: Detecting missing values

The data transformation may be done by any one the following techniques like smoothing, aggregation, generalization, normalization, attribute construction. As for as considering the meteorological data, there may not be any data transformation, as the dataset will be in the appropriate form.

Now the entire dataset is pre-processed and it is ready to perform further process of forming clusters and outlier detection. The pre-processing is done to remove the data with noisy or less deviated data. And then the clustering technique is used to find the outliers. The pseudo code for the Pre-Processing phase is,

Pseudocode

For col = 1 to M
Find mean value "arithmean" of all the attributes of the column „col?
arithmean (col) = add all the elements in the column c/number of elements
    For r=1 to N
        For col = 1 to M
            If D(N,M) is empty(data not yet filled) then
            D(N,M)=arithmean (col)
            End if
    End for
End for

The missing values in the weather data are identified and those data are marked for replacing it with the mean value of that column. So by doing this the missing values are eliminated by filing it with the mean value as shown in the Fig. 5.

sum(values) / values.length

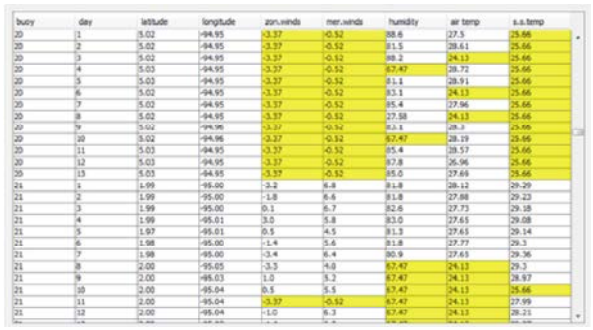

Fig. 5: Filling missing value

Clustering: The grouping of data based on their similarity is known as clustering which is briefly demonstrated in the fig. 7.

In this module the pre-processed data is considered and the further operation such as, cluster and outlier detection is performed. Here the nature of the data is continuous and uncertain; a special methodology is used to track the data. In this case a sliding window technique is used to keep track of the latest data from the stream of available data as shown in the Fig. 6.

In this technique the window size is provided, the window can hold only that many data and those data are called active data. So the process of forming clusters is performed o them. The data within the window will have existential probability; data are clustered based on the distance. The impact of existing data on new element and the impact of new data on existing element are calculated. Then the window now moves on to the next set of data. As the data may influence other data at the time of arrival and also at the time of departure, it is necessary to calculate the probability of the data.

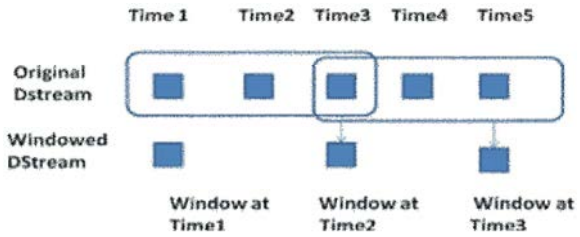

Fig. 6: Sliding Window

And the data that left the window are expired. The same process is repeated. In this way the window is periodically triggered and the clusters are formed for the entire data set.

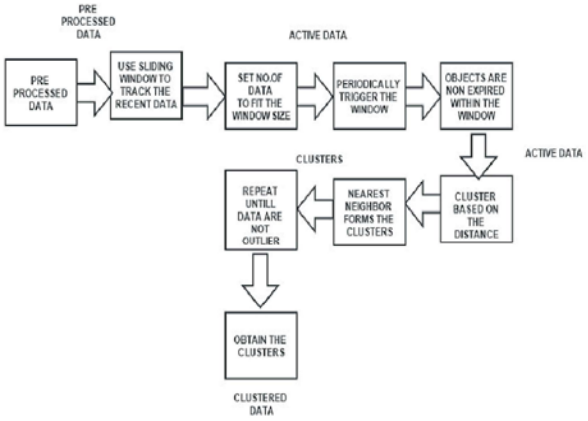

Fig. 7: Clustering

Outlier Detection: The anomaly is removed from the clusters based on the probability as illustrated in the Fig. 8. This helps in forming an efficient data stream which is useful for knowledge discovery.

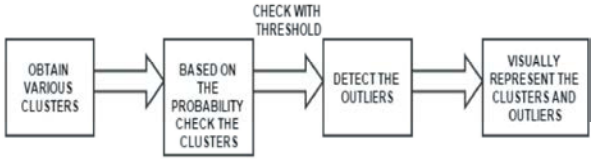

Fig. 8: Outlier Detection

The uncertain data are clustered. And the anomalies are detected are detected based on their probability which can use probability function. Now the data within the clusters are non-outliers. And the visual representation of these clusters shows outliers are removed and only the relevant data are clustered.

This cleaned data can be used for further processing. As the meteorological data is considered, probability has a huge impact. As there are chances of climatic change due to the other data. For example, consider a cluster have been formed based on the temperature like hot, warm, cold.

Now if any is supposed to be absent in a cluster then that cluster would lead to different prediction of weather while classifying. This decision making is shown in the Fig. 9. We can use any of the classification algorithms.

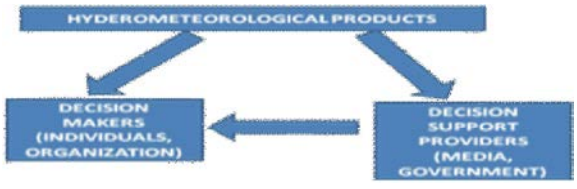

Fig. 9: Weather Prediction

So training of these cleaned data and then classifying them according to the application will be efficient and provides us with the best pattern. So for these purpose the uncertain meteorological data are considered and outliers are removed efficiently.

Clustering and Outlier Detection

DS is the data stream
s is the radius threshold value for clustering
δ can be determined by the available memory

Input: A data stream DS, the radius threshold value s(initial), W- size of the window, ε- error of the window;

Output: The collection of macro clusters MCS.

```
initial MCS = Ø
while DS ≠ 0 do
    read p from DS, generate TC ({ p})
    calculate d (TC ({ p}), MCi ), MCi ? MCS // To check
    probability or to calculate distance
    d min = min{d (TC ({ p}), MCk )}
    if d min ≤ s then //to check if the distance is less than
    the cluster radius
        Mck = MCk + TC ({ p})
        if MCk > β then
            merge the two oldest TC s //forming clusters
            end if;
    Else
            MCS = MCS ∪ TC ({ p})

    if MCS > δ then
            Merge the two nearest MC s
            end if;
    end if;
    DS = DS - { w}
end while;
```

Experimental Results: The experiment is carried out using the well known eclipse tool. There are numerous records in the data set. So the processing times various according to the number of records and the amount of data to be pre-processed. So considering the whole data set, the processing time is,

NO OF RECORDS: 784 ms
PREPROCESSING TIME: 654 ms
MEAN REPLACEMENT TIME: 428 ms
DUPLICATE DETECTION TIME: 14 ms
MISSING VALUE DETECTION TIME: 1 ms

Now to check for the efficiency measure, the data set is divided into two halves. So considering the data set with minimal missing values, the processing time is,

NO OF RECORDS: 385 ms
PREPROCESSING TIME:76 ms
MEAN REPLACEMENT TIME:124 ms
DUPLICATE DETECTION TIME:3 ms
MISSING VALUE DETECTION TIME:1 ms

Even though the records are reduced to half, the processing time varies a lot. This is due to the minimal missing values. So it takes less time to pre-process. If the data with maximal missing values are detected, then the pre-processing time will be increased. From the above experiments using the data set, the processing time for the data is directly proposed to the no of records with missing values.

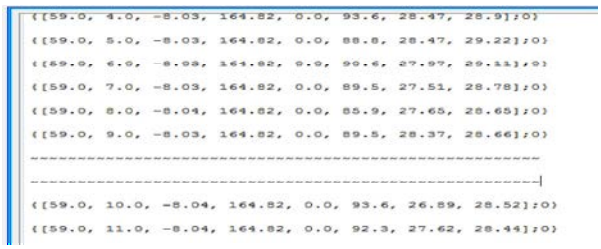

Fig. 10: Elnino clusters

Based on the distance the data in the data set forms different clusters. Considering the elnino data set, there are nearly 800 tuples in the data set. So it is necessary to maintain the enough memory to execute them. The clusters formed by the execution of the program are as shown in the Fig. 10.

The graphical representation of the clu\sters formed is as shown in the Fig. 11. It is clear that from the samples, three clusters are formed. This clustering is based on their nearest neighbors.

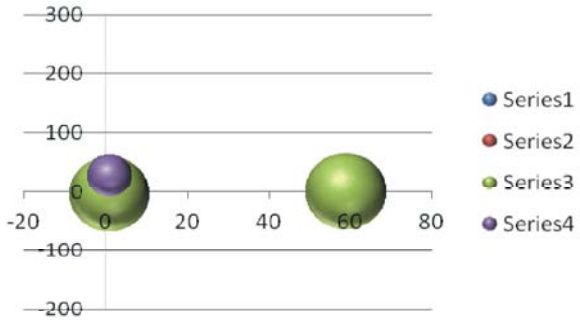

Fig. 11: Clusters

CONCLUSION

The outlier detection in the uncertain data stream having data with underlying probabilities is an extremely crucial component in many data stream like meteorological data. As many sectors like tourism, agriculture, fishery all relies upon the weather condition, the outliers in such data will lead to wrong prediction of weather. As the probability of data may influence other data, outlier removal for the data having existential probability has been proposed. Since the data are uncertain and continuous, the most recent data are tracked using the sliding window. The outliers are detected for the active data based on their probability. And then the cleaned data can be used for knowledge discovery and many other purposes.

REFERENCES

1. Nithya. Jayaprakash, Ms. Caroline Mary, Detection and Deletion of Outliers from Large Datasets. International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 5, July 2014.
2. Ke-Yan Cao, Guo-Ren Wang, Dong-Hong Han, Guo-Hui Ding, Ai-Xia Wang and Ling-Xu Shi, Continuous Outlier Monitoring on Uncertain Data Streams. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(3): 436{448 May 2014.
3. Yufei Tao, Xiaokui Xiao, Shuigeng Zhou, Mining Distance based Outliers from Large Databases in Any Metric Space. ACM, August 20–23, 2006.
4. Tian Zhang, Raghu Ramakrishnan, Miron Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD ACM 1996.
5. Jihyun Ha, Seulgi Seok, Jong-Seok Lee, Precise ranking method for outlier detection. Elsevier 88-107, 2015.
6. Maria Kontaki, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos, Efficient and flexible algorithms for monitoring distance-based outliers over data streams. Published by Elsevier Ltd.,2015.
7. Vijay Babu, K., Dr. R. Shriram, Outlier detection in data streams. International Journal of Future Innovative Science and Technology (IJFIST), Vol(xx) Issue-xx, May 2015..

8. Ali Rahmani, Salim Afra, Omar Zarour, Omar Addam, Negar Koochakzadeh, Keivan Kianmehr, Reda Alhajj, Jon Rokne, Graph-based approach for outlier detection in sequential data and its application on stock market and weather data. Elsevier 2014.
9. Salman A. Shaikh, Hiroyuki Kitagawa, Efficient distance-based outlier detection on uncertain datasets of Gaussian distribution. Springer, 17 April 2013.
10. Prof. Dr. P K Srimani, Malini M Patil, Outlier Mining in Data Streams Using Massive Online Analysis Framework. International Journal of Conceptions on Computing and Information Technology, 3(1).