Outlier Detection in Data Streams Using Fuzzy C-Mean Clustering, Outlier Detection and Genetic Algorithm

Pragya Giri*[1], Abhishek Raghuvanshi[2]
*[1]Research Scholar,MIT Ujaain,M.P,India.
[2]Assistant Professor,MIT Ujjain,M.P,India.
*Department of Computer Science & Engineering.

ABSTRACT: Outliers with the rest of the data points which are different from or inconsistent. New novel, unusual, abnormal or may contain noise. Outliers are sometimes more interesting than the majority of the data. Increasing complexity, size and variety of datasets with major challenges outlier detection, a group, and how to evaluate similar as outliers outliers are to catch. This paper is an approach to detect outlier as a pre-processing step that uses semi surveillance describes outlier detection and then the fuzzy c-means clustering and genetic algorithm to cluster analysis dataset applies to analyze the effects of outliers. As data is digitized, connected and integrated systems, getting the scope of data and analyzes has been growing rapidly. Today, the system's most massive, the size, volume, speed of the phenomenon is changing rapidly, and the non-stationary data generated by these types of data are called data streams. Stream data and your issues in detail in this paper we explore the different techniques for non-reviewed and presented the same results.

Keywords: Outlier Detection, Data Streams, Data Preprocessing, Fuzzy C-mean clustering and genetic algorithm.

I. INTRODUCTION

Now an example of real-time monitoring of data streams in a day for many applications, medical systems, Internet traffic, communication networks, financial market transactions online, remote sensors, and industry are causing the production process. Ordered temporal data streams, rapidly changing, massive, and potentially infinite sequence data objects [1]. Unlike traditional data set, to store a complete data stream or the tremendous amount of time to scan through the impossible. Time data streams can keep evolving new concepts. An evolutionary concept to continually update your model requires data stream processing algorithms to adapt to changes. Data mining is the outlier detection. It is also known as non-mining. Quite a thing for a non-isolated or other data objects is inconsistent. Many applications are more interesting than the usual cases of outliers. Network intrusion detection, credit card fraud detection, weather forecasting, remote detection of cases of medical data, is an example of marketing and customer segmentation.

Frozen outlier detection and data in the literature, based on a specified distance outlier detection algorithm [2, 11], depending on the variety of approaches that are stored in the are, such as the nearest neighbors based outlier detection [3] and in outlier detection based clustering [4, 12]. Most of the data of the current non-detection methods require multiple scanning and / or are a frequent complication. Non-search data streams cannot be used for such methods. Although some data streams [5, 6] Because clustering is based on the detection of non-clustering-based methods are unsupervised in nature, requires knowledge of the data in advance and is not sensitive to low parameters. Relevant and irrelevant characteristics (noise), but equally important ways. The noise property leads to poor performance on real-world data. With the help of the example we will explain on the basis of the drawbacks of existing clustering noise properties outlier detection techniques for data streams.

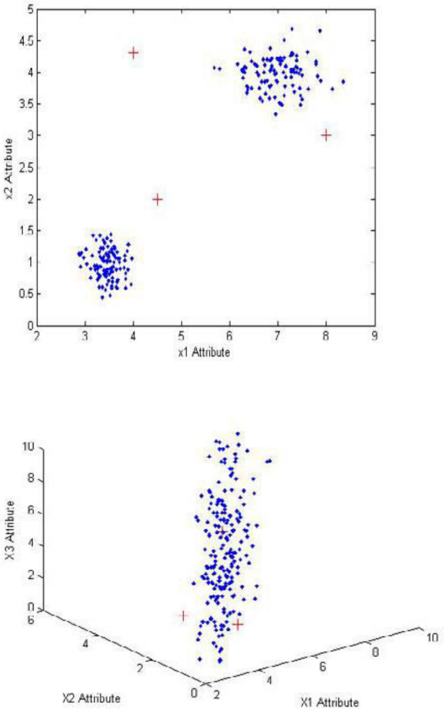

Fig. 1: A scenario showing effect of noisy attributes on clustering and outlier detection. Outliers are represented by red colored '+'symbol and other data objects by blue colored dots (a) Plot of the subspace corresponding to X1, X2 attribute. (b) Plot of the data sets in all dimensions X1, X2 and X3

1.1. MAJOR CHALLENGES AND ISSUES IN OUTLIER DETECTION

Stream data network traffic analysis, sensor networks, such as Internet traffic for different applications, which attributes the irrelevant noise or stream properties which causes challenges in data mining process against it may behave as the system is called might include are produced. There outlier analysis data streams comes from a single data stream or multiple data streams are different issues on the basis of detection of fraud, plagiarism, the communication network is useful in applications such as mining process management. For data stream. In the case of a single data stream issues included below are discussed [2].

Transient: Specific data point to be important for the specific amount of time after it is discarded or stored.

Concept of time: The temporal reference is associated with a timestamp data, based on the temporal reference data point is processed.

The notion of Infinity: The dataset in particular time data stream from the source are produced indefinitely, summary of the available data points are not used.

Arrival rate: Data points comes at different rates, the processing of the data points can be completed before the next data point comes Otherwise, it may result in flooding.

Concept drift: Due to the change in environment, there are changes in the data streams of data distribution Introduced the concept of the characteristics of the data flow is called as.

Uncertainty: The data points can be uncertain external events, are uncertainty factors affecting Ambiguity, vagueness, ambiguity, etc.

Multi-dimensional: Matrix multidimensional data parallelism should be used for the detection of outlier.

II. LITERATURE REVIEW

S.No.	Author Name	Title	Outcome	Limitations

1.	Madhu Shukla, Y.P.Kosta, Prashant Chauhan	Analysis and Evaluation of Outlier Detection Algorithms in Data Streams	1. Different algorithms for outlier detection 2. Outlier detection algorithms that are, weighted attribute method, 3. Neighborhood based algorithms 4. Continuous outlier detection for stream data are given	1. Larger space requirement to store all the window objects.
2.	Yogita, Durga Toshniwal	A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering	1. Evolving data streams that automatically assigns weights to attributes based on their respective significance. 2. Smaller weights reduce the effect of noisy attributes. 3. Proposed framework is incremental and dynamic in nature. 4. This method gives higher outlier detection rate and lower false alarm rate than CORM and LOF. 5. Running time of the proposed method is also less	1. Small sized clusters as outlying clusters are considered. 2. There is a fall in detection rate of proposed method after adding more than 20 % of outlier.

			than LOF.	
3.	Richa Sampat, Shilpa Sonawani	Network Intrusion Detection Using Dynamic Fuzzy C Means Clustering.	1. Based on the combination of feature selection and improved dynamic fuzzy C means algorithm. 2. That improves the performance results of classifiers while using a reduced set of features. 3. The method gives impressive detection accuracy and detection rate. 4. Open-source machine learning and data mining software, WEKA includes many java packages such as associations, classifiers, clusters and so on.	1. It does not implement the traditional fuzzy.
4.	Binu Thomas 1 and Raju G	A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining	1. New fuzzy clustering algorithm combines the positive aspects of both crisp and fuzzy clustering algorithms. 2. It is more efficient in handling the natural data with outlier points than both k-means and fuzzy c-means algorithm. 3. It achieves this by assigning very low	1. The algorithm has limitations in exploring highly structured crisp data which is free from outlier points. 2. The efficiency of the algorithm has to be further tested on a comparatively larger data set.

			membership values to the outlier points.	

III. EXISTING METHOD

K-mean algorithm is used to classify the objects into clusters based on the mean value of the objects in the cluster. K-Mean cluster involves following steps:-

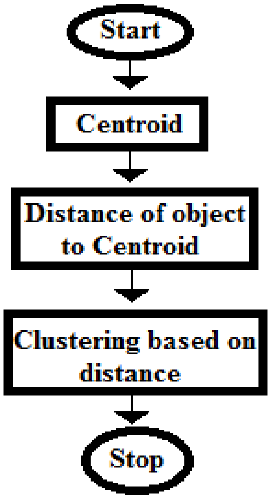

Fig.2. K-Means Algorithm using Clustering.

K-means clustering technique is simple, and we begin with a description of the basic algorithm. We first choose k initial centroids where k is a user-specified parameters, namely, the number of groups desired. Each point is assigned to the nearest centroid, and the centroid of the points assigned to each storage group. Assigned to the cluster centroid of each cluster is updated on the basis of points. Repeat steps until we work and update groups no point change, or rather, until the centroids remain.

8.1 k-means algorithm is described formally. K-means is the operation of Figure 8.3, which shows how, starting from three centroids, is found in the final group stages in four working-update is illustrated. K-means clustering to display these and other figures, each subfigure show (1) Walking and (2) at the start of the work points to those centroids centroids. Centroids "+" symbol are indicated; All points from one cluster size is the same marker.

The basic steps of the algorithm are: -

1: Select points as k initial centroids.
2: repeat
3: By assigning each point to its nearest centroid Form K groups.
4: Each cluster centroid recompute.
5: centroids do not change until.

3.1. Limitations of K-means algorithm

The main limitation of the algorithm for data points in assigning cluster membership comes from its crisp nature. Depending on the shortest distance, the data always becomes a member of one of the groups. It works extremely well with structured data. Real-world data is not arranged in groups around the obvious. Instead, the groups often ill-defined blur the boundaries in the data space around the perimeter of the overlapping groups [4]. In most cases, the real-world data clear external data points clearly does not belong to any of the groups, and they are called non-issue. K-means algorithm and non-overlapping groups dealing with the issue in one of the existing groups to include a data point is not able to do because of this extreme outlier of a cluster of points based on the minimum distance to be covered.

IV. PROPOSED SYSTEMS

In this section detailed description of proposed framework is presented. Pictorial representation of framework is given in Figure 2.

4.1. Preliminary Concepts

Data Stream - A Data Stream DS = x1, x2,......... , XN an infinite sequence of data objects. Object XI = (X1i, x2i,........... , XMI) M is marked by a set of attributes. We share the data stream as the data is processed. N Each part of the data set number of points allowed.

Outlier Detection - Given a data stream DS, a chunk size n, number of clusters k, weight vector w = (w1, w2,......wm) detection of outlier is to find an object which deviates more from its cluster centre than other points in same cluster until L number of chunks.

4.1.1. Data Pre-processing Block

Real-world data sets are highly susceptible to missing or inconsistent data. Such datasets are of low quality and low-quality mining results because the mining quality of the results depends on the quality of data. The block data pre-processing technique is applied to improve the quality of data. Pre-processing techniques, missing values, data scaling, aggregation, general, discretization technique etc. which is applied depends on the type of data stream includes handling.

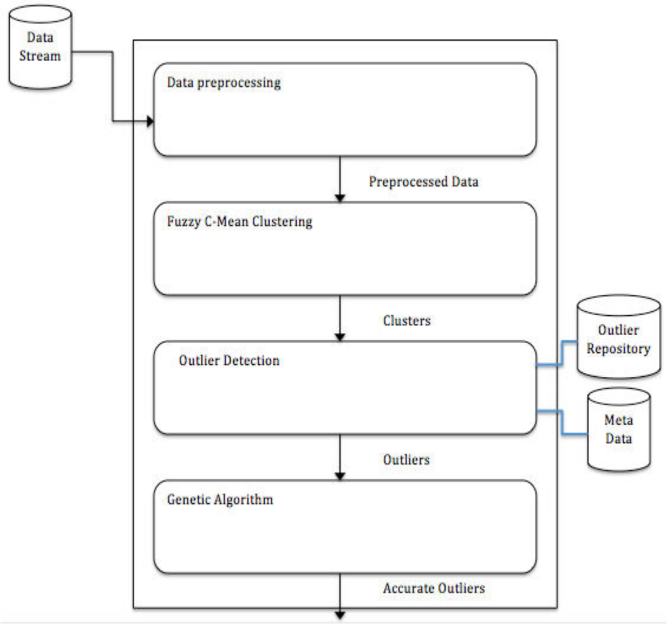

Fig. 3: Proposed Framework for Outlier Detection in Evolving Data Streams by Fuzzy C-Means Clustering.

4.1.2. Fuzzy c-means clustering Cluster centers and the Euclidian distance calculated using a form of the work centers: there are two processes. This process is repeated until the cluster centers stabilize offers included. Fuzzification algorithm parameter m in a range [1, n] need to determine the degree of fuzziness in groups. When M1 reaches the value of the algorithm works like a crisp segmentation algorithm and meters for large values become more overlapping groups. The subscription price with the formula algorithm μ buys kth cluster at any point x (x) is a set of coefficients to be in degree. Fuzzy C-Means, with the centroid of a cluster all points, weighted by their degree of belonging to the cluster means:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}.$$

Algorithm

FCM algorithm is a finite set of data with respect to some given criterion.Given c fuzzy clusters a finite collection of elements in a collection attempts to partition, a partition matrix algorithm and returns a list of cluster centers

Where each element indicates the degree to which element, the cluster belongs to.

FCM is to reduce the objective function:

$$\arg\min_C \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^m \left\|\mathbf{x}_i - \mathbf{c}_j\right\|^2,$$

where:

$$w_{ij}^{m} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|\mathbf{x}_i-\mathbf{c}_j\|}{\|\mathbf{x}_i-\mathbf{c}_k\|}\right)^{\frac{2}{m-1}}}.$$

4.1.3. Genetic Algorithm

A genetic algorithm based search optimization techniques based on the mechanics of natural selection. A genetic algorithm creates a collection of possible solutions for specific problems. Initially the solution randomly generated initial results are so poor, but that our solutions are the solutions desired to see the small segment. The flow of genetic algorithm steps are:

1. Population: = value.
2. Generte_ Kromojom_papuleshn Chromosome_papuleshn = ();
3. The solution then is satisfied if
Sacking
And jump to the next step.
4. Evaluate the fitness value.
5. Number of generation: = value
6. While the number generation * 2 ≤ termination condition; Tax
7. All genetic solutions for the next generation Select Promotions
8. Number_jnreshn ++.
9. foreign operation is to perform 50% of the bits are crossing.
10 is the end.
11. If the solution is efficient
Again
Perform mutation.

Genetic algorithm finds the best solution from a large set of solutions.

V. CONCLUSION

In this paper, we use a fuzzy C- mean algorithm, a new algorithm for outlier detection proposed. The proposed algorithm counts the number of outliers in a particular period of time is good. Future work with a variety of algorithm changes required to implement the proposed work for more dataset for is to make it more efficient. It also proposed to implement a system for distributed environments, processing speed and improve the performance of the algorithm has been planned.

VI. REFERENCES

1. Madhu Shukla, Y.P.Kosta, Prashant Chauhan, “Analysis and Evaluation of Outlier Detection Algorithms in Data Streams”, IEEE International Conference on Computer, Communication and Control (IC4-2015).

2. Yogitaa, Durga Toshniwala, “A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering”, Scienece Direct Procedia Technology 6 (2012) 214 – 222.

3. Richa Sampat, Shilpa Sonawani, “ Network Intrusion Detection Using Dynamic Fuzzy C Means Clustering”, International Journal of New Technologies in Science and Engineering
Vol. 2, Issue. 4, October 2015, ISSN 2349-0780.

4. Binu Thomas and Raju G,” A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining”, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

5. Dr.Chandrika.J, Dr.B.Ramesh , Dr.K.R.Ananda kumar and Raina.D.Cunha, **“**GENETIC ALGORITHM BASED HYBRID APPROACH FOR CLUSTERING TIME”, Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014 pp. 39–52, 2014. © CS & IT-CSCP 2014.

6. Meenakshi Sharma, "Data Mining: A Literature Survey," International Journal of Emerging Research in Management &Technology,2014,pp.1- 4.

7. ShibleeSadik ,LeGruenwald, "Research Issues in Outlier Detection for Data Streams," SIGKDD Explorations Volume 15, Issue 1,2012, pp. 33- 40.

8. Yogita, DurgaToshniwala, "A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering," 2ndInternational Conference on Communication, Computing & Security, pp. 214–222, (ICCCS-2012).

9. Yogita,DurgaToshniwal, "Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering," World Academy of Science, Engineering and Technology, Vol:6, Nov 2012.

10. F. Angiull, F. Fassetti, "Distance-based outlier queries in data streams: the novel task and algorithms," Data Mining and Knowledge Discovery, 20(2),2010, pp. 290–324.

11. DimitriosGeorgiadis, Maria Kontaki, AnastasiosGounaris, Apostolos Papadopoulos, Kostas Tsichlas, YannisManolopoulos, "Continuous