Hybrid Approach for Outlier Detection in High Dimensional Dataset

Rohini Balkrishna Gurav[1], Sonali Rangdale[2]

Student, ME [IT], SCOE, Sudumbare, Pune, India

Guide and ME [IT] coordinator, SCOE, Sudumbare, Pune, India

Abstract: *An object that does not obey the behavior of normal data objects is called as Outlier. In many data analysis process, a large number of data are being recorded or sampled as data set. It is very important in data mining to find rare events, anomalies, exceptions etc. Outlier detection has important applications in many fields in which the data can contain high dimensions. Resulting the intended knowledge of outliers will become inefficient and even infeasible in high dimensional space. I devised an outlier detection structure which is based on clustering. Clustering is an unsupervised type of data mining and it does not require trained or labeled data. Combination of density based and partition clustering method for taking improvement of both densities based and distance based outlier detection. Weights are allocated to attributes depending upon their individual significance in mining task and weights are adaptive in nature. Weighted attributes are useful to reduce or remove the effect of noisy attributes. In view of the challenges of streaming data, the schemes are incremental and adaptive to concept development. In high dimensional data the number of attributes associated with the dataset is very large and it makes the dataset unmanageable. Thus a Feature Extraction technique is used to reduce the number of attributes to a manageable value.*

Keywords: Attribute weighting, Dataset, DBSCAN, k-mean, unsupervised method

1. Introduction

The Object in data set that does not obey to well defined concepts of expected behavior is called Outlier. Outlier detection is preprocessing step for data analysis. In which process of finding objects in the data set that do not follow to particular notions of expected behavior. Detected instances are not behaved like other instances in data set called outliers. It is also called as anomalies or surprises etc. Outlier detection is very essential process for much practical application as E-Commerce; intrusion detection; research etc. Existing methods are classified into 3 categories, supervised, semi-supervised and unsupervised.

To detect outliers in high dimensional data using different clustering techniques. This outlier detection method can be used to find the anomalies in behavior of certain objects in the dataset. This holds importance in the field of Medicine, industries, Network Intrusion etc.

Outlier detection in streaming data is very challenging because streaming data cannot be scanned multiple times and also new concepts may keep evolving in coming data over time. Inappropriate attributes can be termed as noisy attributes and such attributes further enlarge the challenge of working with data streams. The capacity of data in various fields such as medicine, internet transactions is enormous. The outlier detection strategy used for streaming data can be extended for various high dimensional data. Adaptive and dynamic approach can be used for outlier detection in high dimensional space. Detecting outliers in high dimensional data is fast process in data mining. The increasing use of high dimensional data increases the need of finding outliers.

2. Related Work

Outlier detection is very important for data mining research community. Ramaswamy et al proposed a distance based outlier detection method. According to which, given parameters k and n, an object is an outlier ifno more than n-1 other objects in the dataset have higher value for D*k* than object o, where D*k*(o) denotes the distance of k*th* nearest neighbor of object o. This idea is further developed in, where each data point is ranked by the sum of distances from its k*th* nearest neighbors. Breunig et al introduced the notion of the local outlier factor LOF, which captures the relative degree of outlierness of an object. Above described methods are either distance based or nearest neighbors based that are not suitable for outlier detection in data streams due to their high time complexity. He et al in presented new definition of outlier which they named as cluster-based local outlier, which provides importance to the local data behavior. Duan et al proposed a cluster based outlier detection algorithm which can detect both single point outliers and cluster-based outliers. But all these technique that I have defined above and many more are planned for stored static data sets and are not appropriate for data streams environment.

3. Proposed Work

A. Problem statement

Outlier detection in streaming data is very challenging because streaming data cannot be scanned multiple times and also new concepts may keep evolving in coming data over time. Irrelevant attributes can be termed as noisy attributes and such attributes further magnify the challenge of working with data streams

.

B. Basic concepts

- Data Stream - A DataStream is an unbounded sequence of data objects. Object is described by a set of n attributes. I have processed data stream in form of data chunks. Every data chunk contains specified number of n points.
- Data Chunk - Stream of data is an unbounded sequence of data. As it is not possible to store complete data stream, for processing divide it into data chunks of same size. In this paper an object will be observed over multiple successive data chunks before announcing it as outliers
- Weight - Weight of an attribute gives its degree of importance or importance in data stream mining. Large weights are assigned to applicable attribute and smaller too noisy attributes. Weights of all are always one. When we take weights in distance calculation we get good measure of distance.
- Outlier Detection - Given a data stream DS, chunk size l, and weight vector w detection of outlier is to find objects which departs from normal clusters and small in numbers until L number of chunks. These objects can be in collections or individual.
- Variance Matrix - Each entry of this matrix stores sum of distance of all object from their corresponding cluster in corresponding attribute.
- Maximum Score (L) - It specify how many data chunks outlier of an objects is verified before declaring it as real outlier.
- Candidate Outlier - An object is candidate outlier if it differs more than the given threshold from standard clusters based upon deviation measures.
- Real Outlier - A candidate outlier becomes real outlier when it satisfies deviation criteria up to L data chunks.

C. System Architecture

a. Data Chunks & Data Pre-processing -

This process suggests the spreadable division of the data set into data chunks. Processing large number of data at a time is a block hence; I divide data into smaller chunks which can be processed easily without any overhead. Data chunk size can decided by the user.

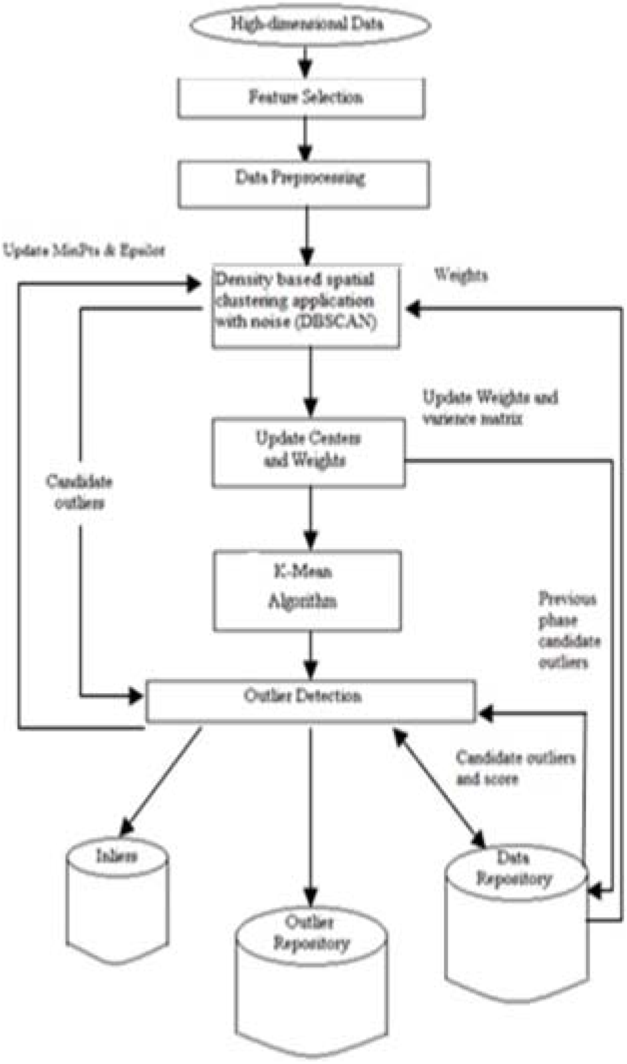

High-dimensional Data
Feature Selection
Data Preprocessing
Update MinPts & Epsilon
Density based spatial clustering application with noise (DBSCAN)
Weights
Update Weights and variance matrix
Update Centers and Weights
Candidate outliers
K-Mean Algorithm
Previous phase candidate outliers
Outlier Detection
Candidate outliers and score
Inliers
Outlier Repository
Data Repository

Figure 1: Proposed work

b. Feature extraction

This stage basically deals with the selection of attributes; wherein the attributes get selected depending on their relevance. The noisy attributes are separated. Thus the attributes is reduced to a count that can be easily processed.

c. DBSCAN algorithm

DBSCAN stands for Density Based Spatial Clustering Application with Noise. It is a Clustering algorithm which forms random shaped clusters. It forms clusters by evaluating minimum points within a minimum distance 'd'.

- **Updation Model:** In this Module the center values are updated.
- **K means algorithm:** This algorithm reduces the data clusters to get an optimized result wherein objects are placed in suitable clusters. This clustering technique uses the mean functionality to calculate mean value as objects get added to cluster. This algorithm provides the guaranteed of the object being allocated to the suitable cluster. Iteratively follow the procedure of cluster pattern until we get the cluster with right objects in it.

d. Outlier detection model

The real outliers are detected in this stage. Candidate outliers are evaluated to discover out if they are real outliers.

4. Results

4.1 User Directory

Figure shows the user directory where dataset are stored. User selects the dataset as per his requirements.

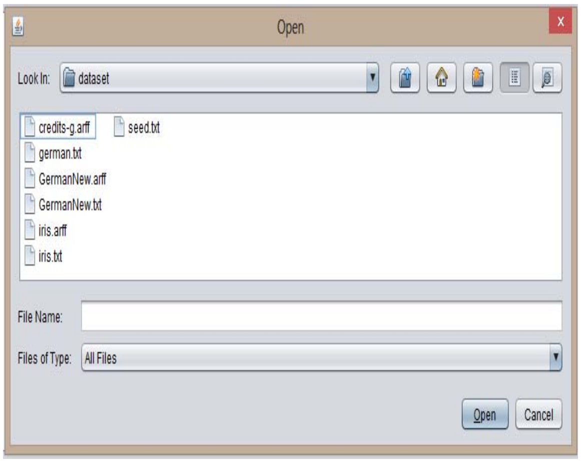
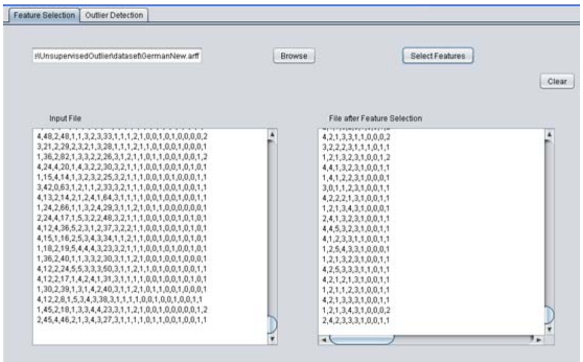


Figure 2: Data set selection

4.2 Feature selection window

Figure shows the output window for feature selection. Here the input window contains all the attributes whereas the feature selection window contains the selected attributes.



```
:\UnsupervisedOutlier\dataset\GermanNew.arff
```

Input File

```
4,48,2,48,1,1,3,2,3,33,1,1,1,2,1,0,0,1,0,1,0,0,0,0,2
3,21,2,29,2,3,2,1,3,28,1,1,1,2,1,1,0,1,0,0,1,0,0,0,1
1,36,2,82,1,3,3,2,2,26,3,1,2,1,1,0,1,1,0,0,1,0,0,1,2
4,24,4,20,1,4,3,2,2,30,3,2,1,1,1,0,0,1,0,0,1,0,1,0,1
1,15,4,14,1,3,2,3,2,25,3,2,1,1,1,0,0,1,0,1,0,0,0,1,1
3,42,0,63,1,2,1,1,2,33,3,2,1,1,1,0,0,1,0,0,1,0,0,1,1
4,13,2,14,2,1,2,4,1,64,3,1,1,1,1,0,0,1,0,0,1,0,0,1,1
1,24,2,66,1,1,3,2,4,29,3,1,1,2,1,0,1,1,0,0,0,0,0,0,1
2,24,4,17,1,5,3,2,2,48,3,2,1,1,1,0,0,1,0,0,1,0,1,0,1
4,12,4,36,5,2,3,1,2,37,3,2,2,1,1,0,0,1,0,0,1,0,1,0,1
4,15,1,16,2,5,3,4,3,34,1,1,2,1,1,0,0,1,0,0,1,0,1,0,1
1,18,2,19,5,4,4,4,3,23,3,2,1,1,1,0,0,1,0,1,0,0,1,0,1
1,36,2,40,1,1,3,3,2,30,3,1,1,2,1,0,0,1,0,0,1,0,0,0,1
4,12,2,24,5,5,3,3,3,50,3,1,1,2,1,1,0,1,0,0,1,0,0,1,1
4,12,2,17,1,4,2,4,1,31,3,1,1,1,1,0,0,1,0,0,1,0,1,0,1
1,30,2,39,1,3,1,4,2,40,3,1,1,2,1,0,1,1,0,0,1,0,0,0,1
4,12,2,8,1,5,3,4,3,38,3,1,1,1,1,0,0,1,0,0,1,0,0,1,1
1,45,2,18,1,3,3,4,4,23,3,1,1,2,1,0,0,1,0,0,0,0,0,1,2
2,45,4,46,2,1,3,4,3,27,3,1,1,1,1,0,1,1,0,0,1,0,0,1,1
```

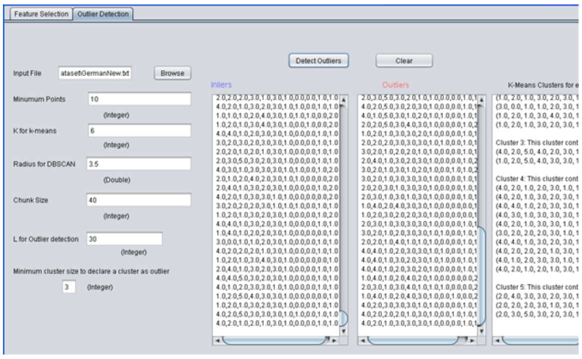
File after Feature Selection

```
4,2,1,3,3,1,1,0,0,0,2
3,2,2,2,3,1,1,1,0,1,1
1,2,1,3,2,3,1,0,0,1,2
4,4,1,3,2,3,1,0,0,1,1
1,4,1,2,2,3,1,0,0,0,1
3,0,1,1,2,3,1,0,0,1,1
4,2,2,2,1,3,1,0,0,1,1
1,2,1,3,4,3,1,0,0,0,1
2,4,1,3,2,3,1,0,0,1,1
4,4,5,3,2,3,1,0,0,1,1
4,1,2,3,3,1,1,0,0,1,1
1,2,5,4,3,3,1,0,0,0,1
1,2,1,3,2,3,1,0,0,1,1
4,2,5,3,3,3,1,1,0,1,1
4,2,1,2,1,3,1,0,0,1,1
1,2,1,1,2,3,1,0,0,1,1
4,2,1,3,3,3,1,0,0,1,1
1,2,1,3,4,3,1,0,0,0,2
2,4,2,3,3,3,1,0,0,1,1
```

Figure 3: Selecting Feature

4.3 Outlier detection

The outlier detection window lists the set of outliers in different data clusters. Fig 10.4 shows the input and clusters formed after applying various clustering algorithm. Fig 10.5 is a graphical representation of the outliers and inliers.



Input File: atasetGermanNew.txt

Minumum Points: 10 (Integer)

K for k-means: 6 (Integer)

Radius for DBSCAN: 3.5 (Double)

Chunk Size: 40 (Integer)

L for Outlier detection: 30 (Integer)

Minimum cluster size to declare a cluster as outlier: 3 (Integer)

Inliers

```
2.0,2.0,2.0,3.0,1.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,2.0,1.0,3.0,2.0,3.0,1.0,1.0,0.0,1.0,1.0
1.0,1.0,1.0,2.0,4.0,3.0,1.0,1.0,1.0,0.0,2.0
1.0,2.0,1.0,3.0,4.0,3.0,1.0,0.0,1.0,0.0,2.0
4.0,4.0,1.0,2.0,3.0,3.0,1.0,0.0,0.0,1.0,1.0
3.0,2.0,3.0,2.0,3.0,3.0,1.0,0.0,0.0,1.0,1.0
2.0,2.0,1.0,2.0,1.0,3.0,1.0,0.0,0.0,1.0,1.0
2.0,3.0,5.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,3.0,1.0,3.0,3.0,3.0,1.0,1.0,0.0,1.0,2.0
2.0,1.0,2.0,4.0,2.0,3.0,1.0,1.0,0.0,0.0,2.0
2.0,4.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,2.0
4.0,2.0,1.0,3.0,3.0,1.0,1.0,0.0,0.0,0.0,2.0
3.0,2.0,2.0,2.0,3.0,1.0,1.0,1.0,0.0,1.0,1.0
1.0,2.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,2.0
4.0,4.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
1.0,4.0,1.0,2.0,2.0,3.0,1.0,0.0,0.0,0.0,1.0
3.0,0.0,1.0,1.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,2.0,2.0,2.0,1.0,3.0,1.0,0.0,0.0,1.0,1.0
1.0,2.0,1.0,3.0,4.0,3.0,1.0,0.0,0.0,0.0,1.0
2.0,4.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,4.0,5.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,1.0,2.0,3.0,3.0,1.0,1.0,0.0,0.0,1.0,1.0
1.0,2.0,5.0,4.0,3.0,3.0,1.0,0.0,0.0,0.0,1.0
1.0,2.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1.0
4.0,2.0,5.0,3.0,3.0,3.0,1.0,1.0,0.0,1.0,1.0
4.0,2.0,1.0,2.0,1.0,3.0,1.0,0.0,0.0,1.0,1.0
```

Outliers

```
2.0,3.0,5.0,3.0,2.0,1.0,1.0,0.0,0.0,1.0,1
4.0,2.0,5.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,1
4.0,1.0,5.0,3.0,1.0,2.0,1.0,1.0,0.0,1.0,1
2.0,2.0,5.0,3.0,4.0,3.0,1.0,0.0,0.0,1.0,2
1.0,2.0,1.0,3.0,3.0,2.0,1.0,0.0,0.0,1.0,1
2.0,2.0,2.0,3.0,3.0,3.0,1.0,1.0,0.0,1.0,1
3.0,2.0,1.0,2.0,2.0,1.0,1.0,0.0,0.0,1.0,1
2.0,4.0,1.0,3.0,2.0,3.0,1.0,0.0,0.0,1.0,2
3.0,2.0,1.0,3.0,1.0,2.0,1.0,0.0,1.0,1.0,2
3.0,2.0,1.0,3.0,3.0,3.0,1.0,0.0,0.0,1.0,1
2.0,2.0,3.0,1.0,3.0,3.0,1.0,0.0,0.0,1.0,1
2.0,3.0,1.0,3.0,1.0,3.0,1.0,0.0,0.0,1.0,1
1.0,4.0,2.0,4.0,3.0,3.0,1.0,0.0,0.0,0.0,1
1.0,2.0,3.0,2.0,2.0,3.0,1.0,0.0,0.0,0.0,1
2.0,3.0,1.0,3.0,3.0,3.0,1.0,0.0,0.0,1.0,1
3.0,2.0,1.0,2.0,3.0,3.0,1.0,1.0,0.0,1.0,1
2.0,2.0,1.0,4.0,1.0,1.0,1.0,0.0,0.0,1.0,1
4.0,4.0,1.0,2.0,3.0,3.0,1.0,1.0,0.0,1.0,1
4.0,4.0,1.0,3.0,2.0,3.0,1.0,1.0,0.0,1.0,1
4.0,4.0,1.0,2.0,3.0,2.0,1.0,0.0,0.0,1.0,1
1.0,4.0,1.0,2.0,4.0,2.0,1.0,0.0,0.0,0.0,2
2.0,3.0,1.0,3.0,4.0,1.0,1.0,1.0,0.0,0.0,2
1.0,4.0,1.0,2.0,4.0,3.0,1.0,0.0,1.0,0.0,2
4.0,2.0,3.0,3.0,3.0,1.0,1.0,0.0,0.0,1.0,1
4.0,2.0,2.0,2.0,1.0,3.0,1.0,0.0,1.0,1.0,1
4.0,2.0,1.0,3.0,3.0,3.0,1.0,0.0,0.0,1.0,1
```

K-Means Clusters for e

```
(1.0, 2.0, 1.0, 3.0, 2.0, 3.0, 1
(3.0, 0.0, 1.0, 1.0, 2.0, 3.0, 1
(1.0, 2.0, 1.0, 3.0, 4.0, 3.0, 1
(1.0, 2.0, 1.0, 3.0, 2.0, 3.0, 1

Cluster 3: This cluster cont
(4.0, 2.0, 5.0, 4.0, 2.0, 3.0, 1
(1.0, 2.0, 5.0, 4.0, 3.0, 3.0, 1

Cluster 4: This cluster cont
(4.0, 2.0, 1.0, 2.0, 3.0, 1.0, 1
(4.0, 2.0, 1.0, 3.0, 2.0, 3.0, 1
(4.0, 4.0, 1.0, 2.0, 3.0, 3.0, 1
(4.0, 3.0, 1.0, 3.0, 3.0, 3.0, 1
(4.0, 2.0, 1.0, 3.0, 3.0, 1.0, 1
(3.0, 2.0, 2.0, 2.0, 3.0, 1.0, 1
(4.0, 4.0, 1.0, 3.0, 2.0, 3.0, 1
(4.0, 2.0, 2.0, 2.0, 1.0, 3.0, 1
(4.0, 1.0, 2.0, 3.0, 3.0, 1.0, 1
(4.0, 2.0, 1.0, 2.0, 1.0, 3.0, 1

Cluster 5: This cluster cont
(2.0, 4.0, 3.0, 3.0, 2.0, 3.0, 1
(2.0, 2.0, 2.0, 3.0, 1.0, 3.0, 1
(2.0, 3.0, 5.0, 3.0, 2.0, 3.0, 1
```

Figure 4: After detecting the outliers

5. Result Graph

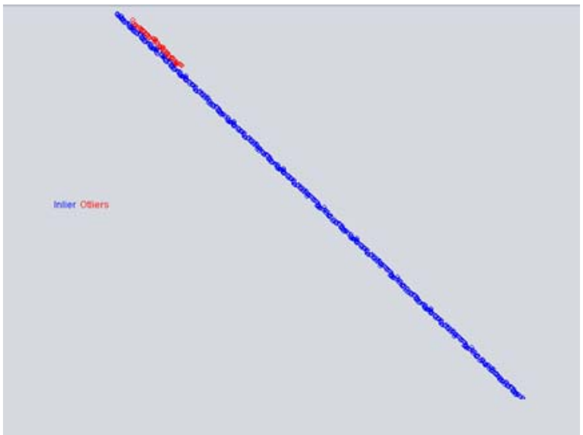


Figure 5: Graph for result

6. Conclusion

In this paper, i have formulated outlier detection as an optimization problem and presents a clustering based unsupervised outlier detection scheme for streaming data. The processed streaming data in the form of data chunks and candidate outliers are checked over multiple consecutive data chunks before declaring them as outliers or inliers. After processing a data chunk only necessary statistics of chunk are kept and chunk is dis- carded to free up memory for next chunk. This scheme has applied both density based (DBSCAN) and partitioning (weighted-k-mean).Clustering for detection of individual as well as group of outliers. With increasing popularity of high dimensional data there increases a need of outlier detection in these datasets. It can be applied to various domains where data with various attributes are handled. This application can be work on mix datasets.

7. Future Scope

The proposed method gives the both rate higher outlier detection rate and lower false alarm which are shown from the experimental result. As comparing with the CORM the performance of the given scheme is very consistent with the increasing number of attributes. In the future we will try to implement advance level of our method for the data types like categorical and mixed.

8. Acknowledgment

With immense pleasure, I am presenting this paper on "Hybrid Approach for Outlier Detection in High Dimensional Dataset". Inspiration and guidance are invaluable in every aspect of life especially in the field of academics, which I have received from respected Head of Information Technology Department Prof. Mr. S.A.Nalawade, PG Coordinator and my guide Mrs. Sonali Rangdale. I would also like to thank all my colleagues who have directly or indirectly guided and helped me in the preparation of this seminar and also for giving me an unending support right from the stage this idea was conceived. I also acknowledge the research work done by all worldwide researchers in this field.

References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.

[2] Yogita and D. Toshniwal, A framework for outlier detection in evolving data streams by weighting attributes in clustering, in Proceedings of the 2nd International Conference on Communication Computing and Security, India, 2012.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, Efficient algorithms for mining outliers from large data sets, in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD 00. New York, NY, USA: ACM, 2000, pp. 427438.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: identifying density-based local outliers, in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD 00. New York, NY, USA: ACM, 2000, pp. 93104.

[5] Z. He, X. Xu, and S. Deng, Discovering cluster based local outliers, Pattern Recognition Letters, vol. 2003, pp. 910, 2003.

[6] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, Efficient clustering-based outlier detec- tion algorithm for dynamic data stream, in Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 05, ser. FSKD 08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 298304.