International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

Grid Partitioning for Anomaly Detection (Gpad) in High Density Distributed Environment for Mining Techniques

C. Viji, N. Rajkumar

Abstract: ***Anomaly detection is the most important task in data mining techniques. This helps to increase the scalability, accuracy and efficiency. During the extraction process, the outsource may damage their original data set and that will be defined as the intrusion. To avoid the intrusion and maintain the anomaly detection in a high densely populated environment is another difficult task. For that purpose, Grid Partitioning for Anomaly Detection (GPAD) has been proposed for high density environment. This technique will detect the outlier using the grid partitioning approach and density based outlier detection scheme. Initially, all the data sets will be split in the grid format. Allocate the equal amount of data points to each grid. Compare the density of each grid to their neighbor grid in a zigzag manner. Based on the response, lesser density grid will be detected as outlier function as well as that grid will be eliminated. This proposed Grid Partitioning for Anomaly Detection (GPAD) has reduced the complexity and increases the accuracy and these will be proven in simulation part.***

Keywords: Grid partitioning, density based outlier detection, Grid Partitioning for Anomaly Detection (GPAD), low complexity, high density environment.

I. INTRODUCTION

Data mining is used to extract the required data from the data sets and that require information will be used for future uses. The data sets are collecting all the data in the database systems. The database system is used to transfer and store the data between the two different users. Databases are split into the different form of sets and then will be formed as data sets. Before the data mining process, it has to apply two different stages such as selection and preprocessing. In the selection stage, the data set will be selected. In the preprocessing stages, noise and unwanted source will be deleted from the dataset. Data mining involves different types of tasks such as:

- Outlier detection
- Clustering method
- Marketing based
- Classification rule
- Regression scheme

Outlier detection is used to detect the unwanted data, data errors and failure rate in the database management system. Clustering method will make the cluster groups and then the task will be done for each clustered group. Marketing based task is based on the dependency modeling scheme. It will do the data mining for repeated datasets. Classification rule will split the data sets into two types one is needed dataset and another as unwanted datasets. Regression scheme is used to find the data model which one has the least error. If any of the datasets have the high error due to the fault, and that will be regretted using the regression scheme.

Outlier detection is the most important task in data mining technology. Here the usage of credit card, outlier issue may generate. The outlier is the outsourced data is trying to mingle to the original source data. That outlier data may increase the failure rate and produces more issue in a form of noises. To avoid that, different approaches are suggested such as proximity-based, model based and cluster based outlier detection.

In proximity based outlier detection, detection of outlier data is based on either distance or density. Compared the distance or density of the data sets are equal to the nearest neighbor data set. If it is equal, the outlier is not affected otherwise, the outlier is detected. In model based outlier detection, implement the data sets into Gaussian model. After the few process, check the Gaussian model status of data sets, if it varies outliers is detected. In cluster based outlier detection, data sets will be split into clustered data sets. Each clustered data sets are compared with their neighbored cluster data sets. The number of data varies, outlier has been detected. To implement this outlier detection in the big data environment, can use the angle based outlier detection scheme. This scheme makes the angel at one point and then another point. Between the two angles, data may vary outlier will be detected.

II. LITERATURE SURVEY

Bai et.al [1] suggests the security mechanism to save the local outlier data from the network intrusion. Here used the two different methods to increase the efficiency such as grid-based portioning algorithm and distributed LOF method. Grid based partitioning algorithm will portion the data into several grid and after that using the second method, each grid will be verified. Ma et.al [2] explains the density based outlier detection algorithm to increase the detection rate in the high-density traffic rates.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
Dr. C. Viji, Associate Professor, Department of Computer Science and Engineering, Akshaya College of Engineering and Technology, Coimbatore, Tamil Nadu, India. (email: vijisvs2012@gmail.com)
Dr. N. Rajkumar, Associate Professor, Department of Computer Science and Engineering, Akshaya College of Engineering and Technology, Coimbatore, Tamil Nadu, India. (email: nrajkumar84@gmail.com)

With the help of spatial-temporal signals, it will avoid the collision and congestion even in high traffic rate. Wang et.al [3] discusses the cluster based outlier scheme to calculate the outlier and to increase the efficiency. This technique uses the cluster approach and it will form the data into clusters. Chang et.al [4] suggests the grid based outlier detection with the help of pruning and searching techniques. This technique will improve the
scalability even in large dense data sources. Gebremeskel et.al [5] suggest the new technique for healthcare safety with the combination of clustered technology and nearest neighbour technique. This technique focuses on to detect the needles for safety purposes in the hospital. Huang et.al [6] explains the natural outlier factor scheme and it will generate the natural values by using the natural neighbour. This technique mainly focuses on the reduction of failure rate and increase the security rate in data mining technologies. Li et.al [7] discusses the grid based outlier detection scheme. It will detect the fraud tolerance in the data mining technologies. This helps to increase the scalability and efficiency of the data mining technology. Zheng et.al [8] proposed the KDE based outlier detection scheme and this scheme helps to applied in the military application by monitoring the environmental surrounding. This technique involves kernel density estimation method and this will update the data after the data transmission between the source and destination node. This technique mainly focuses on the reduction of communication cost and the reduction of fault tolerance rate. Ding et.al [9] explains the hyper grid based anomaly detection algorithm for the searching purposes in the density areas. With the help of real datasets, experiments were conducted and it reduces the region of neighbour detection. Gupta et.al [10] refers the novel based outlier detection scheme to solve the detection of outlier issues. This novel based outlier detection scheme will be applied to the specific datasets for the reduction of failure rate in the data mining. Here this scheme applied to the iris datasets and breast cancer datasets.

III. RESEARCH ISSUES

Distance based outlier detection (DBOD) [11] is used the reverse neighbour count in the unsupervised outlier detection scheme. Here this technique involves the K-NN technique but this will count the neighbours in a reverse manner for the high dimensional data. The normal distance-based outlier detection scheme uses the distance between the neighbours in a straightforward manner due to this process increase the complexity due the data dimension increases. To avoid the difficulty, integrity and complexity, this scheme uses the reverse neighbour counts with an unsupervised scheme for the outlier detection. Continuous angle based outlier detection (CAOD) [12] is used for the high dimension data spaces. The general distance-based outlier detection scheme is not suitable for the high-dimensional data spaces. To avoid that issues, this angle based outlier detection scheme has suggested. By using the incremental algorithm, for every timestamp, this angle based outlier detection process will be implemented. This scheme is used to increase the accuracy even in the high dimensional data environment. Association based outlier detection (ASOD) [13] scheme is used to collect the mixed data and then divide the data sets based on the interval time limit. After the transformation, frequent item set is assigned as the attribute associated. Then assign an order for each data set. Based on the orders of the dataset, an outlier will be detected. Spectral clustering based outlier detection (SCOD) [14] is used to combine the spectral clustering and KNN technique for the efficient performance. This combined spectral technique is used to find the abnormal data in the datasets. Results were compared with the distance based and density based outlier detection. Compared to the other detection schemes, this spectral clustering technique has achieved the high performance. Graph-based Anomaly Detection (GBAD) [15] is a technique involves the data sets are called as graph data. Here implemented the structured graph data for the outlier detection. This technique has compared the performance to the unsupervised and semi-supervised outlier detection scheme. This technique increases the scalability and reduces the robustness using the graph data sets.

IV. PROBLEM FORMULATION

Definition 1: The k nearest neighbor Knn(D) is defined as the kth distance to the point D. And reachability distance between the point D and target point T {R(D,T)} is defined as the max distance between the point D and T {Dst(D,T)} to the k-nearest neighbor point D {Knn(D)}.

This will be defined as,

$R(D,T)=\max_{T}\{\{Dst(D,T),\ Knn(D)\}\}$ (1)

Where D is denoted as data point

T is denoted as target point

Dst is Euclidean distance

Definition 2: Local reachability distance Lr(D) is defined as the ratio of modulo of reachability distance to the summation of reachability distance with respect to the data point D.

$$Lr(D) = \frac{|R(D,T)|}{\sum_{T\in(R(D,T))} R(D,T)} \quad \rightarrow (2)$$

Where |R(D,T)| is having the distance between the target and data point is lesser to the kth nearest neighbor at data point D.

Due to this distance based nearest neighbor technique computational complexity will be higher. This will be overcome by using the density outlier detection approach. That will be updated through local outlier factor.

Definition 3: local outlier factor (LOF) is defined as the inverse of reachability distance with minimum distance with the local reachability distance at data point Lr(D) to the local reachability distance to the target point Lr(T).

$$Lf = \frac{1}{|R(D,T)|}\frac{Lr(D)}{Lr(T)} \quad \rightarrow (3)$$

Due to the different distance and distributed environment, kth nearest neighbor distance based technique may affect the performance and reduce the accuracy rate. This LOF approach is considered based on the density model and this will increase the accuracy rate.

V. GRID PARTITIONING FOR ANOMALY DETECTION (GPAD)

To overcome all those issues and proposed a new evolutionary algorithm for the density based outlier detection is Grid Partitioning for Anomaly detection (GPAD) in the high-density environment. This technique involves in three different stages such as preprocessing, grid partitioning and anomaly detection. Anomaly is the type of data pattern and that behaviour is difficult to predict and it named as outliers, exception, surprise and peculiar. Anomalies are generated in the credit card fraud, intrusion and cyber-crime. Figure 1 shows that the examples of anomalies. Here X1 and X2 are the regions of normal data sets and it is a normal behaviour. The point A1 and A2 are anomalies and it will create the region of A3 and that is also affected by anomalies.

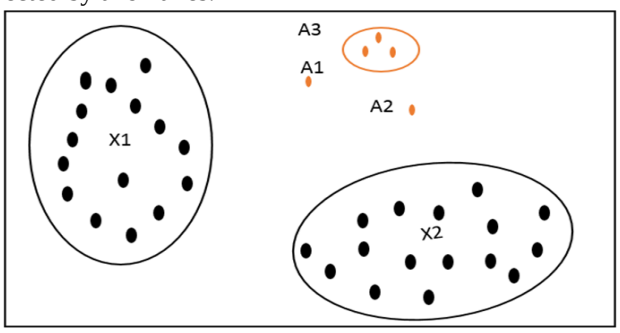

Figure 1. Sample of Density based Anomaly detection

The general process of the proposed Grid Partitioning for Anomaly detection (GPAD) is to detect the anomaly in the large distributed environment. Initially, assign the dataset as D and it contains the data set from d1, d2….dn. these data sets are formed according to the distributed environment. The first step of the proposed Grid Partitioning for Anomaly detection (GPAD) technique divides the data sets in the grid manner. Then the portion of the first grid datasets is questioned to their neighbour dataset for the detection of the outlier. Based on the results, the outlier will be confirmed. The query process will be move on horizontally with a zigzag manner. By using this proposed algorithm can increase the accuracy and transmission quality.

5.1 Preprocessing

The main process of preprocessing stage is for data integration, data cleaning, data transformation and data reduction. Data integration is used to distract the unwanted data from the original data. Data cleaning is used to remove the unwanted noise or error occur in the original data. Data transformation is used to convert the original data into required format for the destination. Data reduction is used to compress the original data without loss of accuracy and quality. Figure 2 shows that the stages of the preprocessing stage. Every dataset are transformed to the preprocessing stage for the outlier detection.

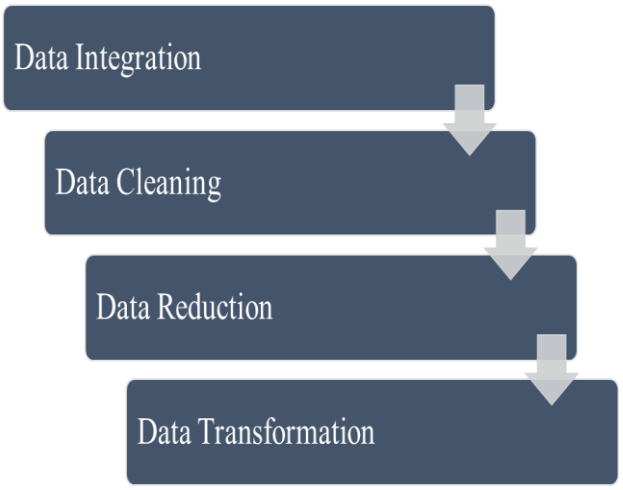

Figure 2. Process of Pre-Processing Stage

5.2 Grid Partitioning

Consider the overall distributed dataset as Ds and that will be splitting into isometric grids. Grid portioning is difficult in high dimensional data space and to avoid that here used the grid portioning algorithm. Initially, data set will be divided into segments and the number of segment is denoted as Ns. And then each segment should be cut into pieces with the grid manner. And the overall distributed data set Ds is splitting into NsDs. If the dimension of the data space is too high, the overall distributed data set Ds is splitting into Ns1.

Let G is Grid and Gx,y is the position of the xth and yth grid. Where, X∈N,Y∈N N denotes the number of grid in the whole data set space.

Assign the data point D to each and every cell of this grid and then compute the density of the ach grid.

If any of the grid have density level is lesser than the particular threshold rate, reject the particular grid. This has reduced the complexity because the complexity will depend only on the number of grids in the whole data space and not in the whole data point in the high dimensional data space.

Theorem 1: Let G is assigned as Grid and it is having maximum number of data points D and Nd is the number of data points available in the each grid. After applying the graph partition, number of data point in each grid will be balanced and that should be lesser than the Nd.

Proof: In each grid has some amount of data points. After applying the graph partition approach, all the D has listed as descending order. And then summing the overall number of data points in the particular dimension space. Take the average of the Nd and that will be denoted as Ag. Now compare the data points of each grid in the zigzag manner to the Ag. If it is lesser and replace the adjacent balancing data load to their respective grid. So after the grid partitioning approach, number of data point will be balanced the load to each grid as well as that value will be lesser than that of Nd. Hence proved.

Figure 3 shows that the Grid portioning approach for the 3X3 matrix.

Y – Dimension	G1,1→ Nd=3	G1,2 → Nd=5	G1,3 → Nd=2
	G2,1 ← Nd=4	G2,2 ← Nd=3	G2,3 ← Nd=4
	G3,1 → Nd=5	G3,2 → Nd=2	G3,3 → Nd=4
	X- Dimension		

Figure 3. Grid Partitioning

Here, Nd is the number of data points in each grid.

Average the number of data point, Ag is 32 = 3.2

After partitioning, all the number of data point is should be greater than the Ag and lesser than the Nd.

5.3 Anomaly Detection

This detection process will start with the query process based on the density level of each grid. By using equation 3, compute LOF for each grid. This will maintain the some threshold rate and this threshold rate is assumption. User can fix the threshold rate based on their application. Threshold rate is denoted as Tr.

For example, from figure 3

G1,1 asks the query to G1,2.

G1,2 response to G1,1. At the same time, G1,2 ask the query to G1,3. And it goes on.

After finishing the first row and it moves to the second row last column.

G1,1 checks the response from G1,2. If the density level of G1,2 is lesser than the threshold rate (Tr). And that grid was affected by anomaly and that grid will be eliminate. Then G1,1 sends the query to G1,3. If the density level of G1,3 is not lesser than the particular threshold rate, and that will be considered as detected Grid. By using the process, outlier detection will be successfully performed. This will help to increase, transmission quality, scalability, accuracy, efficiency and reduces the routing overhead and time cost.

Algorithm:

Input: High density distributed environment

Output: density based outlier detection with grid partitioning

Stage-I: Preprocessing stage

- Assign Data set Ds={D1, D2,……Dn}
- Calculate Dspace; Dimensional Space
- Stage- II: Grid Partitioning
- If Dspace>> λ; λ -Particular level
- G(x,y)=Ns1;
- Else
- G(x,y)= NsDs ;
- End
- For all i=1,2..n; j=1,2..m;
- Assign G(xi,yj) as Dn(k) ; Dn is data node representation for each grid; K=1,..n
- Compute Ng=No. of Ds in Dn(k);
- Compute Ag=∑Ng/100
- If Ng<<Ag
- Insert Ds into Dn(k)
- Else
- Choose Ds in high Dn(k)
- Allocate Ds to G(xi,yj)
- Stage-III: Density based outlier detection
- In G(xi,yj), For r= odd ; \\ r- row for each grid
- Density detection moves towards right
- For r= even; r-row
- Density detection moves towards left
- At the r=z; \\z as end of the row
- Go down
- Move next row r=r+1;
- Compute LOF by using equation 3 in each Grid
- If G(xi,yj) <<Tr
- Eliminate G(xi,Yj)
- Outlier detected
- Else Repeat the process
- End

This algorithm 1 explains the overall process of the proposed Grid Partitioning for Anomaly Detection (GPAD). Grid partitioning helps to reduce the complexity and increases the accuracy. Zig-zag density based outlier detection helps to reduce the routing overhead and time cost.

VI. SIMULATION RESULTS

Simulation results were experimented by using the JAVA programming language. By using the algorithm, program codes were generated. This proposed Grid Partitioning for Anomaly Detection (GPAD) technique is compared with the existing Distance based outlier detection (DBOD) [11], Continuous angle based outlier detection (CAOD) [12], Association based outlier detection (ASOD) [13] and Graph based Anomaly Detection (GBAD) [15]. The parameters are used to compute the performance of these approaches are used as accuracy, scalability, transmission quality, efficiency, routing overhead and time cost.

Accuracy is defined as the ratio of number of corrected predictions to the total number of possible predictions. Transmission quality is defined as the initially generated outsource will be exactly equal to the received outsource and that was received by the destination. Efficiency is defined as the ratio of performance of the outlier detection approach to the percentage. Routing overhead is defined as the ratio of unwanted data sets transmission to the simulation time. Time cost is defined as the required time to detect the outlier in the data grid.

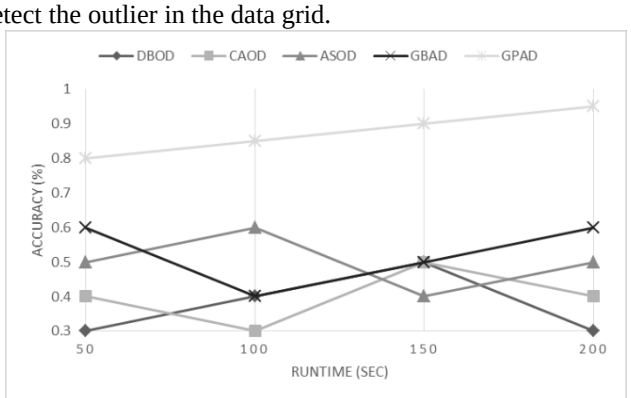

Figure 4. Accuracy Vs Runtime

Figure 4 shows that accuracy with respect to the run time. The run time is assigned as 200 seconds. Between these runtime, the accuracy value will be differ based on the proposed and existing approaches. Compared to the existing approaches such as Distance based outlier detection (DBOD), Continuous angle based outlier detection (CAOD), Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD), the proposed grid partitioning for anomaly detection (GPAD) has increased the accuracy.

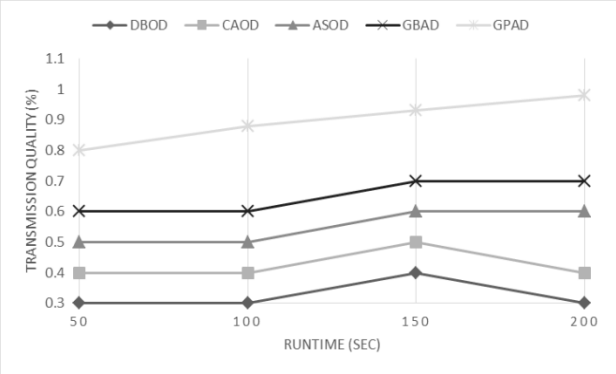

Figure 5. Transmission quality Vs runtime

Figure 5 shows the transmission quality of the proposed grid partitioning for anomaly detection. The existing Distance based outlier detection (DBOD) have lesser quality, Continuous angle based outlier detection (CAOD) have lesser quality, Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD) have lesser quality compared to the proposed technique. This technique helps to increase the quality by increasing the accuracy for outlier detection.

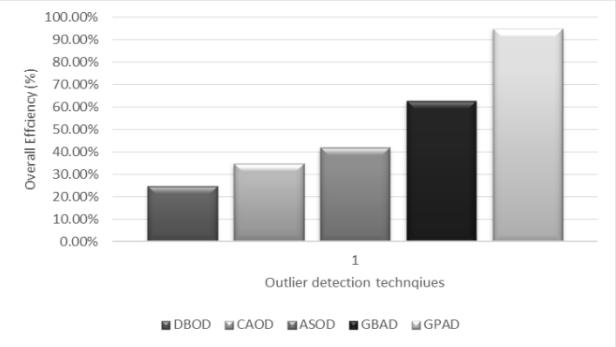

Figure 6. Overall Efficiency Vs Outlier detection techniques

Figure 6 shows the overall efficiency for the existing Distance based outlier detection (DBOD), Continuous angle based outlier detection (CAOD), Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD) and the proposed grid partitioning for anomaly detection (GPAD). The proposed Grid partitioning for anomaly detection increase the efficiency by increasing the accuracy and transmission quality.

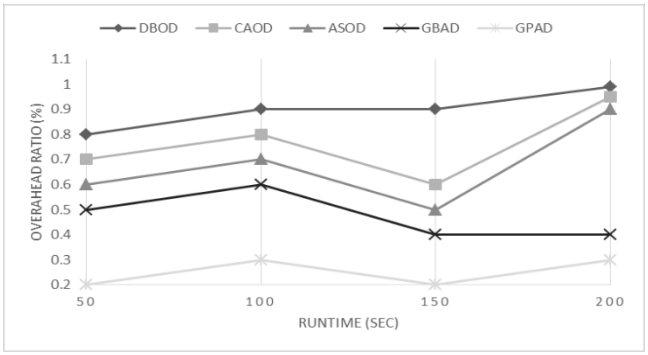

Figure 7. Overhead ratio vs runtime

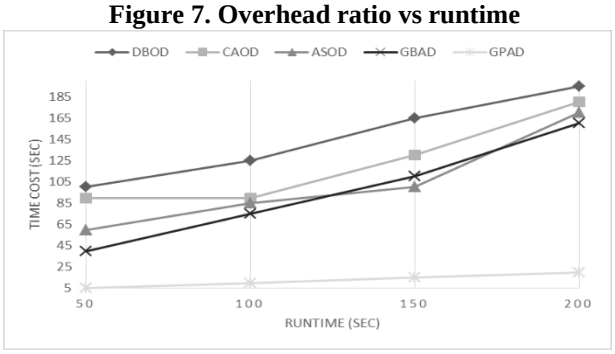

Figure 8. Time cost Vs Runtime

Figure 7 shows the routing overhead versus runtime. The existing Distance based outlier detection (DBOD), Continuous angle based outlier detection (CAOD), Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD) have high routing overhead due to the distance, angle, association and graph respectively. This makes the high routing overhead. But the proposed grid partitioning for anomaly detection (GPAD) has reduced the overhead ratio by dividing the datasets into small subsets with a grid manner.

Figure 8 shows the time cost versus runtime. Distance based outlier detection (DBOD) has high time cost due to the calculation of Euclidean distance to each kth nearest neighbor. Continuous angle based outlier detection (CAOD) has high time cost due to the angle separation and this will not be more efficient one. Association based outlier detection (ASOD) has high time cost due to the steps involved in this process as well high computational complexity. Graph based Anomaly Detection (GBAD) also has high time cost due to the separate the data set in a graphically manner. The proposed grid partitioning for anomaly detection (GPAD) has reduced the complexity due to the grid partitioning and density based outlier detection in a zigzag manner.

Table 1. Comparison of outlier detection techniques with different parameters

Parameter	DBOD	CAOD	ASOD	GBAD	GPAD
Overhead ratio (%)	99	95	90	40	30
Accuracy (%)	30	40	50	60	95
Transmission Quality (%)	30	40	60	70	98
Efficiency (%)	25	35	42	63	95

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 1 shows that the comparison parameters for the existing Distance based outlier detection (DBOD), Continuous angle based outlier detection (CAOD), Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD) and the proposed grid partitioning for anomaly detection (GPAD). Compared to the existing approaches, proposed grid partitioning for anomaly detection (GPAD) has increased the performance of data mining even in high density and high dimensional distributed environment.

VII. CONCLUSION

The proposed grid partitioning for anomaly detection (GPAD) for high density distributed environment and increased the efficiency, accuracy and reduced the routing overhead and time cost. Due to the grid partitioning, can increase the accuracy and this will detect the outlier in each grid. By increasing the accuracy, transmission quality and efficiency will be increased in the grid partitioning for anomaly detection (GPAD) technique. Compared to the existing technique such as Distance based outlier detection (DBOD), Continuous angle based outlier detection (CAOD), Association based outlier detection (ASOD) and Graph based Anomaly Detection (GBAD) and the proposed grid partitioning for anomaly detection (GPAD), the proposed technique has increased the performance. With the help of density based outlier detection in a zigzag manner, helps to avoid the overhead ratio and time cost. In simulation part, by using JAVA programming code, the proposed technique was estimated and that results were prove that the proposed grid partitioning for anomaly detection (GPAD) is suitable for high density distributed environment for data mining techniques.

REFERENCES

1. Bai, Mei, Xite Wang, Junchang Xin, and Guoren Wang. "An efficient algorithm for distributed density-based outlier detection on big data." Neurocomputing 181 (2016): 19-28.
2. Ma, Mathew X., Henry YT Ngan, and Wei Liu. "Density-based Outlier Detection by Local Outlier Factor on Largescale Traffic Data." Electronic Imaging 2016, no. 14 (2016): 1-4.
3. Wang, Xite, Derong Shen, Mei Bai, Tiezheng Nie, Yue Kou, and Ge Yu. "Cluster-Based Outlier Detection Using Unsupervised Extreme Learning Machines." In Proceedings of ELM-2015 Volume 1, pp. 135-146. Springer International Publishing, 2016.
4. Chang, Liang. "GO-PEAS: A Scalable Yet Accurate Grid-Based Outlier Detection Method Using Novel Pruning Searching Techniques." In Artificial Life and Computational Intelligence: Second Australasian Conference, ACALCI 2016, Canberra, ACT, Australia, February 2-5, 2016, Proceedings, vol. 9592, no. 14, p. 125. Springer, 2016.
5. Gebremeskel, Gebeyehu Belay, Chai Yi, Zhongshi He, and Dawit Haile. "Combined data mining techniques based patient data outlier detection for healthcare safety." International Journal of Intelligent Computing and Cybernetics 9, no. 1 (2016): 42-68.
6. Huang, Jinlong, Qingsheng Zhu, Lijun Yang, and Ji Feng. "A non-parameter outlier detection algorithm based on Natural Neighbor." Knowledge-Based Systems 92 (2016): 71-77.
7. Li, Hongzhou, Ji Zhang, Yonglong Luo, Fulong Chen, and Liang Chang. "GO-PEAS: A Scalable Yet Accurate Grid-Based Outlier Detection Method Using Novel Pruning Searching Techniques." In Australasian Conference on Artificial Life and Computational Intelligence, pp. 125-133. Springer International Publishing, 2016.
8. Zheng, Zhigao, Hwa-Young Jeong, Tao Huang, and Jiangbo Shu. "KDE based outlier detection on distributed data streams in multimedia network." Multimedia Tools and Applications (2016): 1-19.
9. Ding, Zhiguo, Minrui Fei, Dajun Du, and Fan Yang. "Streaming data anomaly detection method based on hyper-grid structure and online ensemble learning." Soft Computing (2016): 1-13.
10. Gupta, Raghav, and Kavita Pandey. "Density Based Outlier Detection Technique." In Information Systems Design and Intelligent Applications, pp. 51-58. Springer India, 2016.
11. Radovanovi?, Miloš, Alexandros Nanopoulos, and Mirjana Ivanovi?. "Reverse nearest neighbors in unsupervised distance-based outlier detection." IEEE transactions on knowledge and data engineering 27, no. 5 (2015): 1369-1382.
12. Ye, Hao, Hiroyuki Kitagawa, and Jun Xiao. "Continuous Angle-based Outlier Detection on High-dimensional Data Streams." In Proceedings of the 19th International Database Engineering & Applications Symposium, pp. 162-167. ACM, 2015.
13. Kim, Young-Gi, and Keon Myung Lee. "Association-based outlier detection for mixed data." Indian Journal of Science and Technology 8, no. 25 (2015).
14. Wang, Yuan, Xiaochun Wang, and Xia Li Wang. "A Spectral Clustering Based Outlier Detection Technique." In Machine Learning and Data Mining in Pattern Recognition, pp. 15-27. Springer International Publishing, 2016.
15. Akoglu, Leman, Hanghang Tong, and Danai Koutra. "Graph based anomaly detection and description: a survey." Data Mining and Knowledge Discovery 29, no. 3 (2015): 626-688.