ANOMALY DETECTION OVER DATA SETS - A SURVEY

Sreelakshmi1, Maya Mohan2, and Sruthy Manmadhan3
1M.Tech First Year Student, 2,3Assistant Professor
Department of Computer Science and Engineering,
N.S.S College of Engineering, Palakkad
Email: 1sreelaxmi00@gmail.com, 2mayajeevan@gmail.com, 3sruthym.88@gmail.com

***Abstract*—Anomaly can be considered as something that deviates from what is standard, normal, or expected. data analytics techniques can be used to build the models to detect these problems. Here different techniques are proposed to deal with anomaly detection in data streams using various methods. In the first method mobile network data is utilized as big data which act as a call detail record (CDR) which can help in finding out the anamolus behaviour of the user with the help of methods such as k- means clustering and hierachical clustering. Other methods that can detect and redused anomalies are segment based anomaly detection, sliding window method that focus on angle based process and a method to detect problems in heterogenous data. These anomalies are than compared with the ground truth value to verify the the correctness of the detected anomaly. Here the user activities that were unusually high caused unusual traffic demand and thus were categorized into anomalies. Thus with the help of anomaly detection, region of interests (RoI) can be identified, for which proper action (e.g. proper resource allocation) can be taken in advance to meet the requirements. So by extracting these insightful information, smart and intelligent resource allocation algorithms can be developed for efficient resource utilization.**

***Keywords*: CDR, ROI, Hierarchical Clustering, K-means Clustering.**

I. INTRODUCTION

Anomaly detection is the identication of items, events or observations which do not conform to an expected pattern or other items in a dataset. Anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.Three broad categories of anomaly detection techniques exist:

- Unsupervised anomaly detection
- Supervised anomaly detection
- Semi-supervised anomaly detection

Anomalies in data compromises data quality and can reduce the effectiveness of data effciency. Hence it compromises with the high-throughput real-time analysis of various data streams.So it becomes necessary to detect the anomalies arising in the datas for increasing the effciency and for the increased throughput of any system. Here some of the techniques that can be used to detect anomalies in the data set are discussed such as

- **Anomaly Detection In Mobile Wireless Network**:Here mobile network data (big data) call detail record (CDR) is utilised to analyze anomalous behavior of mobile wireless network using K means clustering and hierarchical clustering.
- **Detecting Insider Threats Using RADISH**:RADISH is a framework which identifies suspicious activity by simultaneously analyzing incoming data streams to learn patterns of normal behavior and, in the context of this learned behavior, search for anomalous activity that portends abnormal system behavior, such as a data breach or attack from within.
- **Sliding Window Based Anomaly Detection**: An angle-based subspace anomaly detection approach is proposed to detect low-dimensional subspace faults from high-dimensional datasets. It selects faulty subspaces by evaluating vectorial angles and computes the local outlierness of an object in its subspace projection.
- **Segment Based Anomaly Detection**: Here a set of neighbouring datasets are considered and these segments are considered as random variables, anomaly is determined by exploiting their spatial predictabilities and, motivated by spatial analysis, specifically investigate how to implement a prediction variance detector in a WSN.
- **Anomaly Detection For Road Traffic**:In this paper a visual analytics framework is proposed that provides support for the anomaly detection in multidimensional road traffic data, the analysis of normal behavioral models built from data, the detection of anomalous events and the explanation of anomalous events.

II. RELATED WORK

A. Anomaly Detection In Mobile Wireless Network

Here mobile network data (big data) call detail record (CDR)[1] is used to analyze anomalous behavior of mobile wireless network. For anomaly detection purposes, a unsupervised clustering techniques is used namely k-means clustering and hierarchical clustering. Then compare the detected anomalies with ground truth information to verify their correctness. The system model

basically consists of LTE-Advanced cellular as shown in Figure.1

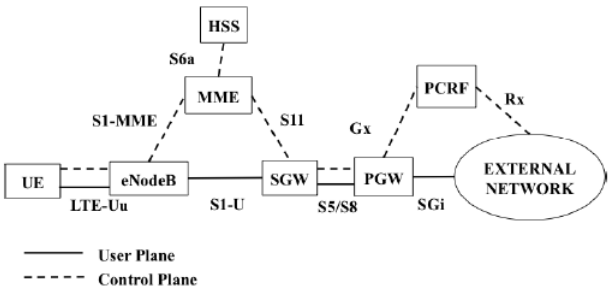

Figure 1. LTE-Advanced Network architecture [1]

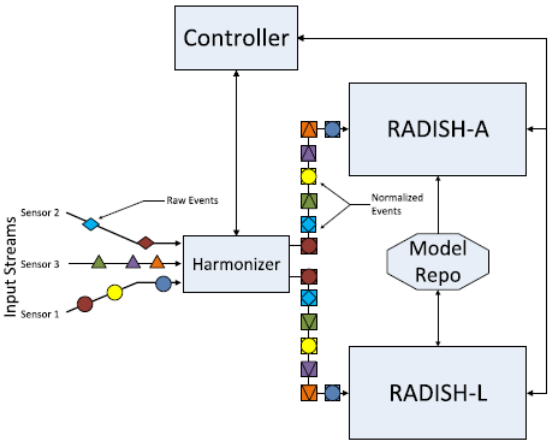

Figure 2. RADISH System Architecture [2]

This technique mainly consist of the following steps:

- Description of dataset:It consist of determining square Id, timestamp, item inbound call activity, item outbound call activity, item inbound sms activity, item outbound sms activity which are aggregated to 1 hour ineterval for the purpose of data preprocessing after the pre processing step the data will be aggregated to a single record.
- Dataset preprocessing:The preprocessing phase consist of cleaning and filtering of data in order to avoid the generation of misleading and inappropriate rules or patterns which constitutes the anomalyand removing the irregularities using Apache pig tool [15] process.
- K-Means clustering [6] algorithm and hierarchical clustering algorithm is used to find out the deviation of the data from the original behaviour and hence calculate the outlierness. Here the number of centroids are determinrd using ELbow method [7].

B. RADISH: Anomaly Detection in Heterogenous Data Stream

This architecture Figure. 2 enables development of models for predictive analytics and anomaly detection as data arrives into the system. RADISH is composed of two distinct processes; a learning process (RADISH-L) and an alerting process (RADISH-A)[2]. RADISH-A matches incoming event streams against the patterns of normalcy derived by RADISH-L to detect anomalous system behavior. An array of sensors is typically deployed within an organization's infrastructure at a variety of places, such as user workstations, routers, firewalls, and servers. The Harmonizer collects the events, transforms them into a common format, performs enrichment and identity resolution, and time-orders the events within a stream.

The main components of the
are:

- **Sensors**: Sensors are devices and applications capable of relaying information on user, asset, or resource utilization to the Harmonizer. Any information source that can communicate over a network can be included as an information stream into RADISH.
- **Harmonizer**: The Harmonizer normalizes events generated by different sensor streams into a common data format in addition to time-ordering them. Harmonizer performs functions like identity resolution, Enrichment, Joining of streams.
- **RADISH-L**: Streaming Machine Learning: RADISH-L ingests events from the Harmonizer, extracts and aggregates relevant features from these events and dynamically creates statistical models representing patterns of normalcy that are then utilized by RADISH-A. RADISH L consist of another step:
  - Online Model Management and Updates:In this step the models built in RADISH A are passed to RADISH L so that it is matched with with the present behaviour to detect the anomaly.
- **RADISH-A: Streaming Anomaly Detection**: RADISH-A uses the models ascertained during the learning phase to identify abnormalities in the event stream coming from the Harmonizer.A multitier architecture is choosen for the RADISH system because it creates a separation between the suspicious activity that is to be found out and the condition which gives rise to alerting situation so as to alert the security team about the anomaly in the system.
- **Controller**:Finally, the RADISH system includes a graphical user interface that is used for the security team to get an overview of the whole system and identify the insider threat attack.. The data come from two main sources from RADISH-A and sensor events from the Harmonizer which is collected from the sensors that are deployed over a large area. These are updated over time, while a context of alerts and behaviors is built up for each user and is displayed in a fashion that allows the analyst to see the correlated behaviors that lead to alerts.

C. Sliding Window Based Anomaly Detection

Here an angle-based subspace anomaly detection approach is proposed to detect low-dimensional subspace faults from high-dimensional datasets. It selects fault-relevant [14] subspaces by

evaluating vectorial angles and computes the local outlier-ness of an object in its subspace projection. It proposes a ABSAD approach Figure. 4 to explore lowdimensional, axis-parallel subspaces that can retain local outlier-ness. This technique is used to find a solution to problem that arise when the concept drifting and the curse of dimensionality is is used simultaneously.

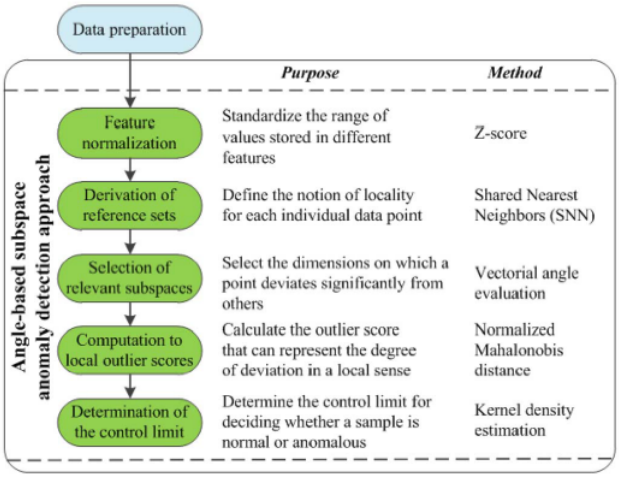

Figure 3. Computational procedure of the ABSAD approach [3]

The various steps used in this methods are as follows:

1) Feature Normalization: The feature normalization step is to standardize the range of values in different features. The Z-score method normalizes the design matrix X to a dimensionless matrix X with zero mean and unit variance. The ith row of X can be calculated as follows:

$$X_i^* = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

2) Derivation of Reference Sets: In low-dimensional spaces, distance-based measures are frequently used to explore the vicinity of a point. SNN approach is the most common approach. The main idea of the SNN method is that two points generated by the same mechanism should have more overlap in their nearest neighbor list, and vice versa. SNN measures the similarity of two points as the number of common nearest neighbors which are derived from a primary measure. Then, the SNN similarity between points p and q can be represented as:

$$Sim_{SNN}(p,q) = \#k_{NN}(p) \cap k_{NN}(q) \tag{2}$$

3) Selection of Relevant Subspaces: The example shown in the following figure will depict how the subspaces are selected in the detection process. Figure. 5 the set RP(p) enclosed by an ellipse contains the nearest neighbors of an outlier candidate p black cross. In the geometrical center of RP(p) is calculated and represented by the point q red circle. Points p and q are connected to form the line l red solid line. In considering which of the 2-D (x and y) p deviates significantly from its reference points, we can evaluate the angle α between line l and the x-axis, and β between line l and the y-axis. The nonsingularity of the covariance matrix relies

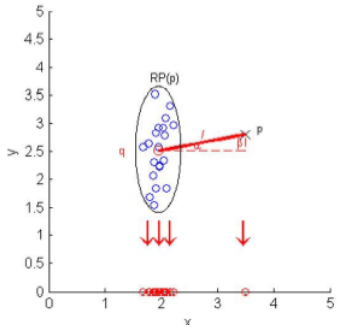

Figure 4. Selection of relevant subspaces [3]

4) Computation to Local Outlier Scores: Normalized Mahalanobis distance to measure the LOS of a specific data point. Mahalanobis distance in the retained subspace normalized by the umber of retained dimensions.

5) Determination of the Control Limit for Reporting Faults: To automate the process of setting control limits, we regard LOS(i), $i\varepsilon 1,2,...,m$ as the observations of a random variable s and apply the KDE method to estimate its probability density function

D. Segment-Based Anomaly Detection

This technique focus handling data in a segment-based manner. Considering a collection of neighbouring data segments as random variables, we determine those behaving abnormally by exploiting their spatial predictabilities and, motivated by spatial analysis, specifically investigate how to implement a prediction variance detector in a WSN. communication cost incurred in aggregating a covariance matrix is optimised using the Spearman's rank correlation coefficient and differential compression, the proposed scheme is able to efficiently detect a wide range of long-term anomalies. Various steps used in the segment based anomaly detection to find the anomaly are:

1) Prediction Varience Detector: In a cluster of nodes Measurement during a period of time are considered as set of ransom variable $Z = X_1, X_2.....X_m$ Any variable $X\varepsilon Z$ can be estimated as linear combination of the remaining variables. Weights in the form of vector can be minimized with help of covarience matrix. It can be solved by lagrange's multiplier and we get a value

$$Var(\varepsilon) = W^T(AW - 2B) + C \tag{3}$$

Detector In a cluster the detector is located at cluster head providing covarience matrix [12]. Varience is employed as the indicator of abnormality. According to Cochrans theorem $Var(\varepsilon)$ can be constructed as a statistical quantity that follows a chi-squared distribution.

2) Approximating Sample Covarience Matrix: Sample covarience matrix is simple unbaised estimator that can be achieved with centralized approach. It is well known that in a WSN the communication cost is often several orders of magnitude higher than the computational cost Wireless communication cost is very high. It will be very expensive for the CH to collect all the local

data segments from the MNs and obtain the sample covariance matrix centrally.

Here a new segment-based anomaly detection technique for handling long-term anomalies by exploiting the spatial correlation existed among neighbouring sensed measurements, with its detector realised through a trackable parameterised statistical quantity, is proposed

E. Anomaly Detection for Road Traffic

analysis of large amounts of multidimensional road traffic data for anomaly detection is a complex task. Visual analytics can be used to bridge the gap between computational and human techniques to detect the anomalies in the multidimensional datas. In this paper a visual analytics framework to detect the anomalies in multidimensional datas which contain the following, the exploration of multidimensional road traffic data, the analysis of normal behavioral models built from data, the detection of anomalous events and the explanation of anomalous events.The anomaly detector implemented is TripMiner as shown in Figure. 6 uses a combination of various data mining methods. There are many methods of anomaly detection for traffic data. A data-driven statistical parametric method, based on the combination of a clustering method (K-means) and a GMM, where the parameters of the Gaussian distribution (mean and covariance matrices) that model the normal data are estimated using the K-means clustering method is used. In order to establish the number of clusters, we used the Silhoutte measurement of closeness for a cluster.

1) Silhoutte estimation and K-means clustering: The silhouette width involves is computed as the between the within cluster tightness and separation from the rest. A clustering can be characterized by the average silhouette width of individual entities; the largest average silhouette width, after testing that giveds different numbers of clusters (K), which indicates the optimal number of clusters [10] in the whole system.

2) Statistical characterization of clusters using GMMs: A GMM is a statistical model in which the overall probability distribution is synthesized from a weighted sum of individual Gaussian distributions. The clustering and the creation of the GMM [8] is done once, using a training data set that represents the normal behavior. When new observations arrive, the GMM can be used to quantify the like $P(d|H = normal), H = hypothesis$

This probability can be calculated with the help of Bayes theorem:

$$P(H = normal|d) = \frac{P(d|H = normal)P(H = normal)}{\sigma_{h\varepsilon H}P(d|h)P(h)} \quad (4)$$

Acceptable ratio can be approximated as :

$$\text{P(H = normalld)} \ \alpha P(d|H = normal)P(H = normal). \quad (5)$$

3) Markov chain and state diagram: Markov chain is a model [9] to determine the future state from the present state provided the conditions remain constant. To account for the temporal variation of the data and the transitions between the different clusters , a Markov chain model is used. Let x_t denote a variable that represents the state of a system at time t (in our case, xt is a multidimensional variable, i.e. $(x_{1t},...,x_{nt})$), where $t = 0,1,2,....$ A stationary Markov chain is a discrete stochastic process.

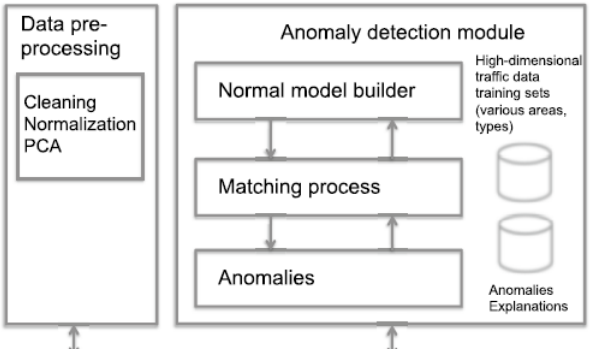

Figure 5. TripMiner framework [5]

III. Discussions and Findings

a) : Anomalies in the dataset compromises with the quality and the efficiency of the data and incorporates unhealthy deviation of the data from original behaviour. Anomaly detetion in mobile wireless network [1] uses k- means clustering algorithm and hierarchical clustering algorithm to find the anomalies in the data set such as call detail record which result in less mean square error than with anomalous data.

b) : RADISH [2] architecture to detect the anomalies in the heterogenous data which contain two parts RADISH-L and RADISH-A in which RADISH-L the learning process makes the learning models of normal behaviour which can be utilized by RADISH-A for alerting the security team to detect the insider threat attacks and fraudlent activities.

c) : Sliding window based anomaly detection mechanism [3] is one which consider the time varying characteristics of a dataset to overcome anomaly problem . This method uses angle based subspace method to detect the anomaly. A segment based anomaly detection technique [4] is used to detect the long term anomalies . TripMiner is a framework proposed in the paper Anomaly detection in road traffic [4] which utilizes visual analytics to find out the anomaly problems. The following can be inferred from these techniques:

While analyzing complexities of K-means clustering and hierarchical clustering for anomaly detection:

- Time complexity of k means: $O(kn)$.
- Time complexity of Hierarchical clustering:$O(n^2 logn)$.
- Space complexity of k-means algorithm is $O(k+n)$.
- Space complexity of hierarchical clustering $O(n^2)$.

So making a hybrid model consisting of k means clustering and hierarchical clustering can balance the time and space complexity. First perform hierarchical clustering on the data sample. The

resulting dendrogram are analysed and decide how many clusters can be formed.Finally perform k means clustering to obtain the anomaly.

No.	Title	Year	Method	Advantages	Disadvantages
1	Anomaly Detection in Mobile Wireless Network	2017	k-means clustering and hierarchical clustering	resulted in less mean square error than with anomalous data	Does not mention about mobility pattern, traffic pattern,energy utilization etc.
2	Anomaly Detection in Heterogenous Data	2017	RADISH-A and RADISH-L	automatically identifies the normal behaviors within an organization	Less capable of alerting detecting insider threat
3	Sliding Window Based Anomaly Detection	2016	angle-based subspace anomaly detection approach	high-dimensional fault detection,online fault detection	Consider less about time varying characteristics of data
4	Segment-Based Anomaly Detection	2015	Segment based method	long-term anomalies are detected	highly dependent on the assumption that the data are spatially correlated.
5	SAnomaly Detection for Road Traffic	2017	TripMiner	allows the analysis of multidimensional data	Increased complexity

TABLE I
COMPARISON

IV. CONCLUSION

Anomaly detection finds out the deviation of an event from its normal behaviour which is identified as anomaly. The solution for various anomalies are found out which are insider threat, financial fraud, and network intrusion is solved by RADISH method, unable to reliably and efficiently report long-term anomalies are solved by segment based anomaly detection and curse of dimensionality and concept rifting is solved by sliding window method.

REFERENCES

[1] Md Salik Parwez, Danda B. Rawat, IEEE and Moses Garuba ,2017,"Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network" IEEE Journals Magazines, Volume: 13, Issue: 4 Pages: 2058 - 2065.

[2] Brock Bse; Bhargav Avasarala; Srikanta Tirthapura; Yung-Yu Chung; Donald Steiner, 2017, "Detecting Insider Threats Using RADISH: A System for Real-Time Anomaly Detection in Heterogeneous Data Streams", IEEE Journals Magazines, Year: 2017, Volume: 11, Issue: 2 Pages: 471 - 482.

[3] Liangwei Zhang; Jing Lin; Ramin Karim, 2017 " Sliding Window-Based Fault Detection From High-Dimensional Data Streams" IEEE Journals Magazines, Year: 2017, Volume: 47, Issue: 2 Pages: 289 - 303.

[4] Grzegorz Blinowski.Miao Xie, Jiankun Hu, and Song Guo,2015, "Segment-Based Anomaly Detection with Approximated Sample Covariance Matrix in Wireless Sensor Networks" IEEE Journals Magazines, Volume: 26, Issue: 2 Pages: 574 - 583.

[5] Maria Riveiro, Mikael Lebram, and Marcus Elmer, 2017,"Anomaly Detection for Road Traffic: A Visual Analytics Framework" ,IEEE Journals Magazines, Volume: 18, Issue: 8 Pages: 2260 - 2270.

[6] Ahmed Zoha, Arsalan Saeed, Ali Imran, Muhammad Ali Imran, and Adnan Abu-Dayya, "A SON solution for sleeping cell detection using low-dimensional embedding of MDT measurements, Mobile Radio Communication (PIMRC), 2014 IEEE 25th Annual International Symposium on, pp. 1626-1630, 2014.

[7] Mingjin Yan, "Methods of determining the number of clusters in a data set and a new clustering criterion," PhD Thesis at Virginia Tech, 2005.

[8] J. B. Kraiman, S. L. Arouh, and M. L. Webb, "Automated anomaly detection processor," Proc. SPIE, vol. 4716, pp. 128 - 137, Jul. 2002.

[9] T. M. Mitchell, Machine Learning. Boston, MA, USA: McGraw-Hill, 1997.

[10] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881 - 892, Jul. 2002.

[11] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," Computer Networks, vol. 38, no. 4, pp. 393-422, 2002.

[12] M.C. Vuran, O.B. Akan, and I.F. Akyildiz, "Spatio-Temporal Correlation: Theory and Applications for Wireless Sensor Networks", Computer Networks, vol. 45, no. 3, pp. 245-259, 2004.

[13] M. E. Houle, H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in Scientific and Statistical Database Management. Berlin Heidelberg, Germany: Springer, 2010, pp. 482500.

[14] M. E. Houle, H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in Scientific and Statistical Database Management. Berlin Heidelberg, Germany: Springer, 2010, pp. 482 - 500.

[15] Online, https://pig.apache.org.