ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.3

A Study on Outlier Detection for Temporal Data

Alavilli Anusha[1], I. Srinivas Rao[2]
M. Tech Student[1], Assistant Professor[2]
Department of Computer Science and Engineering
Thandra Paparaya Engineering College, Andhra Pradesh, India

Abstract:
Data mining provides a way for finding hidden and useful knowledge from the large amount of data .usually we find any information by finding normal trends or distribution of data .But sometimes rare event or data object may provide information which is very interesting to us .Outlier detection is one of the task of data mining .It finds abnormal data point or sequence hidden in the dataset. Data stream is unbounded sequence of data with explicit or implicit temporal context .Data stream is uncertain and dynamic in nature. In the statistics community, outlier detection for time series data has been studied for decades. Recently, with advances in hardware and software technology, there has been a large body of work on temporal outlier detection from a computational perspective within the computer science community. Outlier detection alludes to task of distinguishing examples. They don't accommodate set up normal conduct. Anomaly detection in high-dimensional information presents different difficulties coming about because of the "scourge of dimensionality". The present perspective is that separation fixation that is propensity of separations in high-dimensional information to end up in perceivable making separation based strategies mark all focuses as similarly great exceptions. This paper gives proof by exhibiting the separation based strategy can create additionally contrasting exception in high dimensional setting. The high dimensional can have an alternate effect, by re-evaluating the idea of opposite closest neighbors. In particular, advances in hardware technology have enabled the availability of various forms of temporal data collection mechanisms, and advances in software technology have enabled a variety of data management mechanisms. This has fuelled the growth of different kinds of data sets such as data streams, spatio-temporal data, distributed streams, temporal networks, and time series data, generated by a multitude of applications. There arises a need for an organized and detailed study of the work done in the area of outlier detection with respect to such temporal datasets. In this survey, we provide a comprehensive and structured overview of a large set of interesting outlier definitions for various forms of temporal data, novel techniques, and application scenarios in which specific definitions and techniques have been widely used.

Keywords: Data mining, Outliers, data stream mining.

I. INTRODUCTION

Detection of outliers in information characterized as discovering examples in information that don't comply with ordinary conduct or information that don't fit in with expected conduct, such an information are called as outliers, peculiarities, special cases. Irregularity and Outlier have comparable importance. The experts have solid enthusiasm for outliers since they may speak to basic and significant data in different spaces, for example, interruption detection, misrepresentation detection, and medicinal and wellbeing analysis. An Outlier is a perception in information occurrences which is not quite the same as the others in dataset. There are numerous reasons because of outliers emerge like poor information quality, breaking down of gear, ex charge card extortion. Information Labels associated with information occurrences demonstrates whether that case has a place with ordinary information or atypical. Based on the accessibility of marks for information case, the inconsistency detection procedures work in one of the three models are

1) Supervised Anomaly Detection, systems prepared in regulated mode consider that the accessibility of named occurrences for typical as well as peculiarity classes in a preparation dataset.
2) Semi-managed Anomaly Detection, strategies prepared in directed mode consider that the accessibility of marked cases for typical; don't require lab ls for the oddity class.
3) Unsupervised Anomaly Detection, procedures that work in unsupervised mode do not require preparing information. There are different techniques for outlier detection based on closest neighbors, which consider that outliers show up a long way from their closest neighbors. Such techniques base on a separation or closeness measure to seek the neighbors, with Euclidean separation. Numerous neighbor-based strategies incorporate characterizing the outlier score of a point as the separation to its kth nearest neighbor (k-NN technique), a few strategies that decide the score of a point as indicated by its relative thickness, since the separation to the kth closest neighbor for a given information point can be seen as an assessment of the backwards thickness around it. Outlier detection is the technique which distinguishing Patterns that don't fit in with set up standard conduct. Hawkins characterizes "the outlier as perception that goes amiss to vast degree from the other perception which implies that the example is produced by the distinctive system" Outlier detection is the procedure of discovering information from extensive and multidimensional databases to take in the startling example and conduct of articles. The paper applies the OD on the k-dimensional dataset with k>=5. This methodology utilizes the separation based outlier detection for multidimensional dataset. Bunching is procedure of a gathering of information into gatherings concerning a separation or comparability measure. In information mining, grouping is utilized to disclosure of the conveyance of information and the detection of examples. Here creators have proposed another bunching calculation called C2P. This methodology abuses record structures along the handling of nearest combine questions in spatial databases. It consolidates the upsides of the progressive agglomerative and chart theoretic bunching calculations. The paper gives

augmentation to substantial spatial databases and for outlier taking care of the outlier detection methods work in one of the three modes are;

1) Supervised outlier Detection:
These techniques are trained in supervised mode and consider the availability of labeled instances for normal as well as outlier classes in a training dataset.
2) Semi-supervised outlier Detection:
This technique is trained in supervised mode and considers the availability of labeled instances for normal and do not require labels for the outlier class.
3) Unsupervised Outlier Detection:
These techniques operate in unsupervised mode do not require training data from any class. There are many more outlier detection techniques based on the nearest neighbor which considers that outlier object appears far from their nearest neighbor. Such methods base on a distance or similarity measure to search the neighbors with Euclidean distance. Numbers of neighbor-based OD methods include defining the outlier score of a point as the distance to its kth nearest neighbor.

II. TYPES OF OUTLIER DETECTION TECHNIQUES

2.1 Statistical outlier detection
Statistical outlier detection techniques make assumption about normal data and outlier data. [1] It assumes some data distribution in the data set. Outliers are points that have a low probability to be generated by the overall distribution. [6] A good domain specific knowledge is required. Statistical outlier detection methods are divided into two categories: Parametric methods assume the distribution model priory. [1] Statistical outlier Detection methods use training data set to build the statistical model. In data stream we cannot assume data distribution because we cannot have entire data set and data distribution may change over time. So same training dataset or model cannot produce true outliers. In Non-parametric methods, the model of normal data is learned from the input data. Non-parametric statistical methods does not make any assumption about the distribution of the data so it can be used in data stream with single dimension or very low dimensions [7]. They cannot be applied in high dimensional data stream.

2.2 Distance Based Outlier Detection
Distance based outlier detection technique decides the outlines of the data point based on its distances to its nearest neighbors. Definition [Knorr and Ng]: "Given parameters k and R, an object is a distance-based outlier if less than k objects in the input data set lie within distance R from it."[11] They are defined for any data type for which distance measure or similarity measure is available and these methods do not require detailed understanding of application domain.[10] In distance based methods k-nearest neighbor distance from the original data points are considered for calculating outlier score instead of pre aggregated data so outlier detection is performed at a finer granularity than other methods like Clustering or density based methods .So they can distinguish between noise and true outlier.[3] Distance based methods do not assume any data distribution. So they can be used in data stream. Definition for data stream: "The outlier score of a data point is defined in terms of its k-nearest neighbor distance to data points in a time window of length W."[3] It is not effective for high dimensional data stream. High dimensional dataset in real application contains very much noise. Distance between all data points are equal in high dimensional data point .So degree of outlierness are same for all data point. [7]

2.3 Density Based Outlier Detection
In density based outlier detection, density around a data point is compared with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. Basic assumption in density based outlier detection method: The density around a normal data point is similar to the density around its neighbors. The density around an outlier is considerably different to the density around its neighbors. In this type of method, outliers are detected by computing Local Outlier Factor (LOF), which is the ratio of local density of the point and the local density of its nearest neighbor. Data point whose LOF value is high is declared as outlier. In [13], an incremental LOF algorithm which is suitable density based algorithm for data stream, is proposed. It can detect changes in the data behavior. It provides performance equivalent to static LOF algorithm. It cannot distinguish between outliers and new data behaviors.
In [14], an improvement of Incremental LOF algorithm is proposed .It can significantly distinct outliers from new data behavior. The density based outlier detection methods are more effective than distance based methods [7]. But they are more complicated and computationally expensive because they involve density of both the point and its neighbors also. Density based methods are not effective for high dimensional data set because the accuracy of the density estimation process degrades with increasing dimensionality.

2.4 Sliding window based outlier detection
Streaming data uses sliding window concept which is used for maintaining the statistical information in data stream. The window is identified by two sliding end points. [11] Both ends are active. During the moving process, both ends are moving in the same direction and shifting the same units. Let W be the window size so only the last W records to arrive in data stream are relevant at any point of time. It has some overlaps between the next window and the last window. Here W is fixed based on the number of records or the interval of time. [15] If some data points are outliers in one window, they can be inliers in other window because nature of data stream is dynamic and data behavior may change during the time. Hence detecting any changes in a data as outlier is not desirable. So determining out lierness of a data point as it arrives although meaningful can lead us to wrong decisions. Choosing accurate window size in sliding window based outlier detection is required. Choice of sliding window is independent of data point used for implementation which gives poor result over outlier detection. [6]

2.5 Clustering based outlier detection
Clustering based outlier detection is an unsupervised outlier detection technique in which class label as "normal" or "outliers" are not available. For this reason, clustering means learning by observation rather than learning by examples. [1]Clustering method is used to group similar data points in a cluster. The main requirements for clustering evolving data streams are Summarization, Processing, and Outlier detection. [17] Here we assume that Normal data instances belong to a cluster in the data, while anomalies either do not belong to any cluster, Normal data instances are close to the closest cluster centroid, while anomalies are far from their closest cluster centroid, Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse cluster. [19]. In [18] Clustering based outlier detection

technique for evolving data stream is proposed that assigns weights to attributes according to its relevance in mining task. The testing phase for clustering based techniques is fast because the number of clusters against which every test instance needs to be compared is small. [19] The main aim of most of the clustering algorithms is to find clusters rather than outliers and they are not optimized to find anomalies. Most of the existing clustering algorithm requires number of clusters in advance and shape of the clusters are also defined, but in data stream we cannot assume no of clusters in advance. Arbitrary shape clusters also cause some difficulties in realizing exact clusters of the data.

III. RELATED WORK

3.1. Outlier detection

Definition: "An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism "[1] In many applications, data are generated by processes which may reflect either the activity of the system or observation of objects in the system.[2] Outliers may appear in the data due to some reasons like Mechanical fault, Changes in System behavior, Fraudulent Behavior, Human Error, Instrument error, Natural deviations in populations, changes of environment etc. Outlier can be noise or interesting item. There is no clear distinction between outlier and noise. It depends on interest of the analyst of the system. [2] In both the cases outlier detection is crucial because most of the statistical methods cannot work well in the presence of outliers. [6].Applications of outlier detection are credit Card fraud detection, Telecom fraud detection, Loan application processing to detect fraudulent applications or potentially problematical customers, Intrusion detection detecting unauthorized access in computer networks. Detecting unexpected entries in databases for data mining to detect errors, frauds or valid but unexpected entries.

Types of Outliers:

Global outliers: Global outliers are also called point outliers. A data point is called global outlier if it is different or far from the whole dataset. It is very simple form of outlier. Most of the techniques are designed for finding this type of outliers. For example, Intrusion detection, if a communication behavior of a computer is very different from the normal behavior trends then it is found to be a global outlier. [1] A data stream cannot have a global outlier because it has temporal context available with it. Data stream is infinite and online detection has to be done so only a subset of the data set is available for a specific time instead of the whole data. [9]

Contextual outliers: It is also called conditional Outlier. A data point is called contextual outlier if it is different or far from the other data points in the specific context. Contextual outlier detection techniques provides flexibility for detecting outliers in different contexts. Whether a 30^0 C temperature is outlier or not depends on time and location. If it is winter in Toronto then it is an outlier. If it is summer then it is normal.[1] In contextual outlier detection data attributes are divided into two groups Contextual Attributes: Attributes with respect to which a data point is considered an outlier are contextual attributes. It defines context of the object .For example time, spatial attribute (Longitde and latitude), network location etc.[5] Behavioral attributes: They are non-contextual attributes and evaluated to find outlierness of data point in context in which it belongs. [1] For example Temperature, humidity, pressure, rain fall etc. [5]

Collective outliers: A collection of data point as a whole is different from the entire data set is called collective outlier. An individual data point in a collection may not be outlier. Usually data points are related in collection. Finding subsequence as anomaly in time series data set, finding sub regions as anomaly in spatial imaginary data set or finding sub graph as anomaly in graph data set are examples of collective outliers [5]

Available online: http://internationaljournalofresearch.org

3.2. Data Stream

Today many data sources like sensor network , world wide web, Social networking sites , Telecommunication , Internet traffic, online transactions, medical systems ,Real time surveillance systems and other dynamic processes generate tremendous amount of data every time .They are coming continuously . We cannot store those data in limited memory of our computer for processing or analyzing. They are called stream data. Data streams are uncertain, Dynamic and infinite sequence of data points. [9] Data streams can be of two types Time series data stream: In time series, temporal component is stronger than multidimensional data stream. [3]Time series data stream is treated as contextual dataset where time components can be a contextual attribute. Choice of similarity function is very crucial in time series data analysis. Dimensions in a time series data stream can be defined in two ways depending on the application: In multivariate time series, all behavioral attributes are considered as dimensions and in univariate time series, all values are considered as dimensions. Multidimensional data stream: Multi-dimensional data stream analysis is same as the multi-dimensional static data set analysis but temporal component is added in data stream.[8] Temporal component in multidimensional data stream is weaker than time series data stream .All attributes of multidimensional data stream are treated equally. For outlier detection, time series data require the analysis of each series as a unit, whereas the multidimensional data requires the analysis of each multidimensional point as a unit.

3.3. Outlier Detection in Data Stream

In many cases, the detection of unusual events needs to be performed in a time critical manner so techniques of detecting outliers in data stream must be developed .Most of the algorithms developed are for static dataset which can be stored in memory. Data stream cannot be stored in memory and algorithms have to be applied as it arrives. Currently, most of the existing outlier detection algorithms only put focus on real-time outlier detection in data stream, but ignore subsequent changes of data stream, which means that these algorithms cannot find the mutual conversion between outliers and normal data points. [16] Output of outlier detection algorithms can be

Outlier score: outlier score is assigned to data point according to its degree of outlierness. [5] Data point which has higher value of the outlier score has higher probability to be an outlier. [2] Output of such outlier detection techniques is ranked list of outliers. An analyst may choose to either analyze top few outliers or use a cut-off threshold to select the outliers.

Binary Label: These type of techniques assign a label indicating whether or not a data point is an outlier .This type of output contains less information than the first one because a threshold may be applied on the outlier score to convert them into binary labels. [2]

IV. ISSUES OF OUTLIER DETRECTION IN DATA STREAM

Issues related outlier detection in data stream with respect to data stream characteristics:

4.1 Transient: Data point in data stream are transient in nature .So after some time it loses its importance because it is

discarded or archived. Earlier outlier detection techniques construct the outlier detection model using entire dataset and then data point is compared to the model or other data points to detect whether it is outlier or not .For data stream, outlier detection

4.2. Notion of time stamp

Each data point in data stream is associated with some notion of time implicit or explicit. In explicit association, time is an attribute and in implicit association, exact time is not important but order of the data points is important .If we consider temporal context, a data point is considered outlier if it deviates significantly from the other data points with the same temporal context .an appropriate temporal context has to be decided first and then every data point has to be processed according to its temporal context.

4.3 Infinite

Data stream is an infinite sequence of data points as they keep coming from a data source indefinitely. So at any specific time the entire dataset cannot be available so many static outlier detection techniques which require whole dataset for detecting outlier cannot be used. Outlier detection method for data stream should store summary of the data set and summary should be computed incrementally. A data point is compared with the summary of the data points instead of other data points. Thus an outlier detection model has to be incremental and cannot assume the availability of the entire dataset. [9]

4.4. Arrival rate

Arrival rate may be fixed or variable. Outlier detection technique for data stream has to process data point before next data point arrives .The set of data points or the summary of the data points, to which the current data point is compared to detect outlier-ness, should be adjusted based on the available processing time.[9]

4.5. Concept Drift

The distribution of data may change over time in data stream and it is called concept drift. Data point which has detected as outlier for one data distribution may change its outlierness with changing data distributions. So outlier detection technique for data stream cannot assume any fixed distribution of data. [12]

4.6. Uncertainty

Today new hardware technology such as sensors are generating large amount of data. But data contain missing, inconsistent or erroneous values .Uncertainty indicates it is impossible to determine whether the information available is true or not. Uncertainty in data stream is a key challenge for outlier detection. [10]

V. CONCLUSION

A large number of techniques have been developed in outlier detection area, but most of them have been some inherent limitations. Outlier detection over streaming data was an important research problem in data mining community. Finding out outlier is important because it contains useful information which may lead for further researching domain. This Project provides a review of outlier detection methods over streaming data with data mining perspective. Based on review we can conclude that the most of the techniques used are focuses over algorithms. These require a special background and notion of finding anomaly also varies from domain to domain. It is observed that efficiency of outlier detection method is highly based on data distribution and type of data. Some techniques mentioned in this Project require a prior knowledge about data. For instance that statistical technique uses a data distribution and model. Also the assumption made about data is correct. The individual methods are not efficient over streaming data. In such case if prior information about data is not known then better to make use of combine approach for outlier detection.

V. REFERENCES

[1]. C. C. Aggarwal, Outlier Analysis. New York, NY, USA: Springer, 2013.

[2]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 15:1–15:58, 2009.

[3]. V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.

[4]. Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," Centre Telemat. Inform. Technol. Univ. Twente, Enschede, The Netherlands, Tech. Rep. TR-CTIT-08-59, Oct. 2008.

[5].C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," SIGMOD Rec., vol. 30, pp. 37–46, May 2001.

[6]. C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in Proc. 2008 SIAM Int. Conf. SDM, pp. 483–493.

[7]. C. Aggarwal and K. Subbian, "Event detection in social streams," in Proc. 12th SIAM Int. Conf. SDM, 2012, pp. 624–635.

[8]. C. C. Aggarwal, "On abnormality detection in spuriously populated data streams," in Proc. 2005 SIAM Int. Conf. SDM, pp. 80–91.

[9]. C. C. Aggarwal, Y. Zhao, and P. S. Yu, "Outlier detection in graph streams," in Proc. 27th ICDE, Hannover, Germany, 2011, pp. 399–409.

[10]. J. Gao et al., "On community outliers and their efficient detection in information networks," in Proc. 16th ACM Int. Conf. KDD, 2010, pp. 813–822.

[11] A. Ghoting, M. E. Otey, and S. Parthasarathy, "LOADED: Link based outlier and anomaly detection in evolving data sets," in Proc. 4th IEEE ICDM, 2004, pp. 387–390.

[12]. M. Gupta, J. Gao, Y. Sun, and J. Han, "Community trend outlier detection using soft temporal pattern mining," in Proc. ECML PKDD, Bristol, U.K., 2012, pp. 692–708.

[13]. M. Gupta, J. Gao, Y. Sun, and J. Han, "Integrating community matching and outlier detection for mining evolutionary community outliers," in Proc. 18th ACM Int. Conf. KDD, Beijing, China, 2012, pp. 859–867.

[14]. J. P. Burman and M. C. Otto, "Census bureau research project: Outliers in time series," Bureau of the Census, SRD Res. Rep. CENSUS/SRD/RR-88114, May 1988.

[15]. A. J. Fox, “Outliers in time series,” J. Roy. Statist. Soc. B Methodol., vol. 34, no. 3, pp. 350–363, 1972.

[16]. H. Cho, Y. jin Kim, H. J. Jung, S.-W. Lee, and J. W. Lee, “OutlierD: An R package for outlier detection using quantileregression on mass spectrometry data,” Bio informatics, vol. 24, no. 6, pp. 882–884, 2008.

[17]. V. Barnett and T. Lewis, Outliers in Statistical Data. New York, NY, USA: Wiley, 1978.

[18]. D. M. Hawkins, Identification of Outliers. London, U.K.: Chapman and Hall, 1980.

[19]. P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. New York, NY, USA: Wiley, 1987.

[20]. T. Lane et al., “Sequence matching and learning in anomaly detection for computer security,” in Proc. AAAI Workshop AI Approaches Fraud Detection Risk Manage., 1997, pp. 43–49.

[21]. S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov, “Anomaly detection in large sets of high-dimensional symbol sequences,” NASA Ames Res. Center, Mountain View, CA, USA, Tech. Rep. NASA TM-2006-214553, 2006.

[22]. S. Budalakoti, A. N. Srivastava, and M. E. Otey, “Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety,” IEEE Trans. Syst., Man, Cybern., C, Appl. Rev., vol. 39, no. 1, pp. 101–113, Jan. 2009.

[23]. V. Chandola, V. Mithal, and V. Kumar, “A comparative evaluation of anomaly detection techniques for sequence data,” in Proc. 2008 8th IEEE ICDM, Pisa, Italy, pp. 743–748.

[24]. K. Sequeira and M. Zaki, “ADMIT: Anomaly-based data mining for intrusions,” in Proc. 8th ACM Int. Conf. KDD, New York, NY, USA, 2002, pp. 386–395.

[25]. A. Nairac et al., “A system for the analysis of jet engine vibration data,” Integr. Comput. Aided Eng., vol. 6, no. 1, pp. 53– 66, Jan. 1999.

[26]. X. Pan, J. Tan, S. Kavulya, R. Gandhi, and P. Narasimhan, “Ganesha: BlackBox diagnosis of mapReduce systems,” SIGMETRICS Perform. Eval. Rev., vol. 37, no. 3, pp. 8–13, Jan. 2010.

[27]. U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, “Finding anomalous periodic time series,” J. Mach. Learn., vol. 74, no. 3, pp. 281–313, Mar. 2009.

[28].L. Portnoy, E. Eskin, and S. Stolfo, “Intrusion detection with unlabeled data using clustering,” in Proc. ACM CSS Workshop DMSA, 2001, pp. 5–8.

[29]. M. Gupta, A. B. Sharma, H. Chen, and G. Jiang, “Context-aware time series anomaly detection for complex systems,” in Proc. SDM Workshop, 2013.

[30]. E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, “A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data,” in Proc. Appl. Data Mining Comput. Security, 2002.

[31]. P. Evangelista, P. Bonnisone, M. Embrechts, and B. Szymanski, “Fuzzy ROC curves for the 1 class SVM: Application to intrusion detection,” in Proc. 13th Eur. Symp. Artif. Neural Netw., 2005, pp. 345–350.

[32].J. Ma and S. Perkins, “Time-series novelty detection using one-class support vector machines,” in Proc. IJCNN, Jul. 2003, pp. 1741–1745.

[33]. B. Szymanski and Y. Zhang, “Recursive data mining for masquerade detection and author identification,” in Proc. 5th Annu. IEEE SMC Inform. Assur. Workshop, 2004, pp. 424–431.

[34]. F. A. González and D. Dasgupta, “Anomaly detection using real-valued negative selection,” Genet. Program. Evolvable Mach., vol. 4, no. 4, pp. 383–403, Dec. 2003.

[35].C. Marceau, “Characterizing the behavior of a program using multiple-length n-grams,” in Proc. 2000 NSPW, pp. 101–110.

[36]. C. C. Michael and A. Ghosh, “Two state-based approaches to program-based anomaly detection,” in Proc. 16th ACSAC, New Orleans, LA, USA, 2000, pp. 21–30.

[37]. S. Salvador and P. Chan, “Learning states and rules for detecting anomalies in time series,” Appl. Intell., vol. 23, no. 3, pp. 241–255, Dec. 2005.

[38]. N. Ye, “A Markov chain model of temporal behavior for anomaly detection,” in Proc. 2000 IEEE SMC Inform. Assur. Security Workshop, vol. 166. pp. 171–174.

[39] J. Yang and W. Wang, “CLUSEQ: Efficient and effective sequence clustering,” in Proc. 19th ICDE, 2003, pp. 101–112.

[40]. P. Sun, S. Chawla, and B. Arunasalam, “Mining for outliers in sequential databases,” in Proc. 6th SIAM Int. Conf. SDM, 2006, pp. 94–105.

[41]. E. Eskin, W. Lee, and S. Stolfo, “Modeling system calls for intrusion detection with dynamic window sizes,” in Proc. DISCEX, vol. 1. 2001, pp. 165–175.

[42]. W. Lee, S. J. Stolfo, and P. K. Chan, “Learning patterns from unix process execution traces for intrusion detection,” in Proc.AAAI Workshop AI Approaches Fraud Detection Risk Manage., 1997, pp. 50–56.