A Review on Various Outlier Detection Techniques and Challenges

Er. Divya Sharma[1], Er. Ajay Sharma[2]
[1]*M.Tech Scholar,*
[2]*Associate Professor*
Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab
[1]*div2711@gmail.com*, [2]*ajaysharma@acetamritsar.org*

Abstract- Data mining is an extremely researched area in the today's world as data is critical part of many applications, due to which many researchers express their interest in this domain. As there arises a need to procedure large data set which imposes dissimilar challenges for researchers. To have a data which is free from a noisy attributes, known as a filtered data, is of much importance to increase accuracy in a result set. For that, finding &eliminate the noisy objects has gained a much more importance. An object that does not follow the footprints of the usual data object is called outliers. Outlier detection procedure is used in various applications like fraud detection, intrusion detection system, tracking environmental activities, healthcare diagnosis this survey includes the existing outlier techniques and applications where the noisy data exists. Our paper describes critical review on various techniques used in different applications of outlier detection that are to be researched further & they gives a specific type of knowledge based data i.e. more useful in research activities. So where the Anomalies is present it will be detected through outlier detection techniques & monitored accordingly particularly in educational Data Mining.

Keywords- Point Outliers, Contextual Outliers, Collective Outliers, Intrusion Detection, Fraud Detection, Text data Detection

I. INTRODUCTION

Data mining is a process of extracting hidden and useful information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality. Finding outliers is an important task in data mining.Outlier detection is currently an active and important research problem facing by the many data mining researchers & involved in number of applications. Due to time different nature of the incoming data; declaring an outlier often can lead us to a wrong conclusion. But However, earlier research done over a mentioned problematic of outlier detection is more suitable for static data sets where the whole dataset is readily available and algorithms can operates over multiple passes. But, outlier detection over dynamic data set is actual challenging task because data is always updated and flowing. Finding outliers from a collection of data is a very well-known problem in the domain of data mining [1]. An object that does not follow the footprints of the usual data object is called outliers. In other words, outlier is a pattern which is not similar with respect to the rest of the designs in the dataset. Dependending on the application domain, outliers are of particular interest. Detecting outliers may lead to the discovery of truly unexpected.

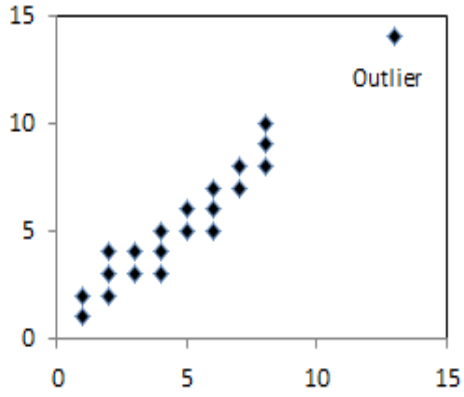

Fig. 1: Outlier detection

behavior and help avoid wrong conclusions etc. Which gives us a filtered and cleared data to operate on. In particular, distance-based techniques use the distance function for relating each pair of objects of the data set. Distance-based definitions represent an useful tool for data analysis [2].

The main focus the outlier detection of Outlier in a Static data set as well as in and continuously data stream which gives us filtered data with which we can carry out experiments on application consisting a huge data set .

II. RELATED WORK

Rajendraet.al (2011). [3] author discussed a clustering based technique to capture outliers. Here they, apply K-means clustering algorithm to divide the data set into clusters. The

points which are lying near the centroid of the cluster are notlikely candidates for outlier & they can pruned out such points from each cluster. Based on the outlier score obtained, we declare the top n points with the highest score as outliers. The experimental results using actual data set demonstrate that smooth though the number of computations is less, the proposed method performs better than the existing outlier detection methods. PrashantChauhanet.al (2015) [4] Author discussed, Various approaches used to achieve the mentioned goal, Some of them use KMeans algorithm for outlier detection in data streams which help to make a similar group or cluster of data points. So they are so-called cluster based outlier detection. Purpose of this paper is to review of different approaches of outlier detection which is used for K-Means algorithm for clustering dataset with some other m.Liu et al. (2004) [5] present an outlier-resistant data filter-cleaner based on the earlier work of Martin & Thomson (Martin & Thomson, 1982). The proposed data filter-cleaner includes an on-line outlier-resistant estimate of the process model and combines it with a modified Kalman filter to detect & "clean" outliers. The future method does not require an apriori knowledge of the process model. It detects and replaces outliers on-line while preserving all extra information in the data. The authors demonstrated that the planned filter-cleaner is efficient in outlier detection and data cleaning for auto-correlated and even non-stationary process data.

III. TYPES OF OUTLIERS

A very important aspect of an outlier detection technique is the nature of the desired outlier. Outlier Classification is complete on the basis of their occurrence; general there are 3 types of outliers which are itemized as follows:

- Point Outliers
- Contextual Outliers
- Collective Outliers.

A. Point Outlier:

When a data instance is different from set of data then instance is called as point outlier. It is the simplest form of outlier & used in various researches. For example credit card fraud Detection, the outlier can be detected through respect to amount spent if expenditure is greater compared to normal transactions then it is an outlier.

B. Contextual Outlier:

When a data instance is anomalous with respect to some context (situation), then instance said to be Contextual Outlier. Contextual outliers generally explored on time series data. For example, in context of age a six feet adult may be a normal person while 6 feet child is an outlier.

C. Collective Outlier:

When a group of related data is anomalous from rest of the complete data set, then it is a collective Outlier. They can occur only in data sets where data instances are related. Collective outlier has been explored on graphical data, sequential data and spatial data.

IV. MAJOR CHALLENGES AND ISSUES IN OUTLIER DETECTION

Stream data are produced from the different applications like network traffic analysis, sensor network, internet traffic etc., which may contain attributes that are irrelevant called as noisy attributes which causes challenges in stream data mining process or it may be anomalistic behavior of the system. Outlier analysis is useful in applications like fraud detection, plagiarism, communication network management. For the data stream mining process there are various issues based on the data streams which comes from the single data stream or multiple data streams. In case of single data stream issues involved are discussed below[6].

A. Transient: Specific data point is important for specific amount of time, after it is discarded or archived.

B. Notion of time: Timestamp attached with data which give temporal context, based on that temporal context data point is processed.

C. Notion of infinity: Data stream are produced indefinitely from the source thus at particular time whole dataset is not available so summary of data points are used.

D. Arrival rate: Data points arrives at the different rate, so processing of data points can be completed before the next data point arrives otherwise, it results in flooding.

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

F. Uncertainty: Due to external events data points may become uncertain, factors affecting are uncertainty, imprecision, vagueness, ambiguity etc.

G. Multi-dimensionality: For outlier detection in multidimensional data similarity matrix should be used.

V. OUTLIER DETECTION METHOD

Outlier detection is used in several domains of applications. It can easily be used with data, image, & software. Basically anomaly detection & misuse is used for removing the noisy data & producing accurate data set. Various applications of outlier detection are enumerated below:

1) Intrusion Detection:

Intrusion detection identifies all of the suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system.

Generally two approaches are used to detect computer security intrusion system in real time: Misuse and Anomaly detection. Misuse detection aims to detect recognized attacks against computer system although anomaly detection uses knowledge of users normal behavior to detect attempted attacks.

Some outlier detection methods used in Intrusion Detection assumed below in Table 1.

Table 1: Intrusion Detection

Name of technique	Type of Intrusion	References
Neural Networks	Network based intrusion, Process based intrusion, Software based intrusion, Geodetic Network (ADALINE algorithm)	[7] - [8], [9]
Statistical Profiling Using Histograms	Stack and Hostbased, Host based, Hostbased, Networkbased, Hostbased and Stack-based	[10] - [11]
Rule Based Systems	Host and Network based	[12]
Parametric Statistical Modeling	System based	[13]
Non-Parametric Statistical Modeling	Network based	[14]

2) Fraud Detection:

Fraud detection is at alarming rate & hence becomes aexcessive threaten for the institution and banks using a credit card transactions Fraud is reported under crime activities that contains banks, mobile phones fraud detection, commercial etc. Outlier is mainly used to detect a noisy data that is being presented in the original data. The various techniques are applied to detect a fraud these are presented in table 2 enumerated below.

Table 2: Fraud Detection

Name of technique	Type of Fraud	References
Neural Network	Credit Card Banking, financial accounting fraud detection	[15, 16, 17, 18]
Novelty Detection	Online short Detection	[19]
Rule-based	Credit card Banking	[20]
Clustering	Credit card Banking	[21]
Bayesian Classification and decision tree	Financial accounting fraud detection	

Software developed and implemented using neural network on Mellons bank mainframe computers resulting in reduced fraud consistently accurate and timeless of fraud detection [15]. The neural network requires high diagnostic quality, to overcome this an algorithm developed that generalizes the transaction data and obtain higher level of diagnostic rules then conflating with rule based information and a classification results in better fraud detection[16].

3) Medical and public health outlier detection

The patient data is to be collected from the several features of patient like blood test, height, weight, patient age. The several methods of outlier detection is used in medical diagnoses which helps to detect critical diseases at early stage for preventing it to become a severe and lifetaking disease. Outlier detection plays a major role in detecting various kinds of cancers. Some outlier detection methods used in Medical & public health are illustrated below in Table 3.

Table 3: Medical and public health

Name of Technique	Type of Disease	References
Parametric Statistical Modeling	ECG	[22]
Bayesian Approach	MYELOMA CANCER	[23]
Cluster based	Lympography and Breast cancer	[24]
Fuzzy Logic	Heart Diseases	[25]
TANAGRA (Data mining tool)	Breast Cancer	[26]
Neural Network	Hypoglycemia	[27]
k-Nearest Neighbor Classifier	electroencephalogram (EEG)	[24]

4) Image Detection

Images can be of any type main purpose of outlier detection is to identify an abnormal behavior of the images. Each data consist of the various features of the image that includes color, brightness, image co-ordinates and texture. The techniques to detect an outlier in images are illustrated in table 4

Table 4: Image Detection

Name of Technique	Features of image	Reference
Regression	Wavelength	[28]
Clustering	Hyperspectralimage	[29]
Neural Networks	Sequential images	[30]
Mixture of Models	Mammogram	[31]
Classification	Spatiotemporal data in image sequence	[32]
Support Vector Machines	Multispectral and hyperspectral images using segmentation	[33]
Hidden Markov Model	General images	[34]

5) Text data Detection

Noisy data is current in the pile of contents that is to be detected through the outlier methods. The data can be spatial or can be a temporal means spatial related to the geographical conditions & temporal related to the time aspects. The actual aim of outlier detection is to handle the noisy data that is presented in the pile of text. Various methods for the detecting anomalies in Text are enumerated in Table 5.

Table 5: Text Data

MultiScale Approach	Spatial and Temporal Data	[35]
Statistical Approach	Multivariate Data[56], Text data[39]	[36], [37]
Clustering, Kgeneration, Thompson's Tau method, maxflow min-cut algorithm.	Multidimensional Data	[38]
Local Outlier Factor	Synthetic and Real time data	[39]

6) Sensor Networks

Now a days sensor networks are used in the many applications of day to day life activities. A sensor network is a combination of specialized transducers with ability of communication which helps to monitor and record situations like humidity, pressure, vibrations, intensity of sound, level of pollution&concentration of chemicals etc. at different locations. A sensor network is a communication system which intends to record conditions & monitor at many locations. A sensor network have multiple detection station called sensor node. Each node is portable, less weighted and very minor in size. Basically through the outlier detection methods faulty sensor networks are detected so that communication level is to be increased. Reliability in wireless sensor networks is artificial by the various causes like atmosphere condition, using low quality sensors etc. that leads to corrupted data generations by sensor containing missing values. Specific outlier detection techniques used in Sensor Net-works Table 6.

Table 6: Sensor Networks

Name of technique	Kinds of networks	References
Bayesian Networks	Wireless sensor networks	[40]
Rule-based	Wireless sensor networks	[41]
Nearest Neighborhood Based Techniques	Multisensor network Event based, multisensor network(car and light sensor)	[42, 43, 44]

VI. OUTLIER TECHNIQUES

The outliers detection techniques debated in this study are:

A. Statistical: Statistical techniques fit a statistical model, usually for normal behavior, to the given data and then a statistical inference test is applied to determine if an unseen instance belongs to the model or not. Instances that have a low probability to be generated from the learnt model, based on the applied test statistic, are declared as outliers [45].

B. Clustering: Is used to group similar data into clusters. Even though clustering & outliers detections appear to be basically altered from each other, some clustering based outlier detection techniques have been developed.

C. Classification: This technique is used to learn a model from a set of labeled data instances and, then, classify a test instance into one of the classes using the learned model. Classification based anomaly detection techniques operate in a similar two-phase: the training phase learns a classifier using the available labeled training data and the testing phase classifies a test instance as normal or anomalous using the classifier [45].

D. Nearest Neighbor: Require a distance or similarity measure defined between two data instances, that can be computed in different ways [45]. Techniques based on this method can be broadly grouped into 2 categories: techniques that use the distance of a data instance to its nearest neighbor as the anomaly score and techniques that compute the relative density of each data instance to compute its anomaly score.

E. Mixture Models: Mixture models comprise a finite or infinite number of components, possibly of different distributional types, that can describe different features of data [46]. In statistics, a mixture model is a probabilistic model for density estimation using a mixture distribution. A combination model can be regarded as a type of unverified learning or clustering.

F. Spectral: Try to find an approximation of the data using a combination of attributes that capture the bulk of variability in the data [45]. This technique defines subspaces in which the anomalous instances can be easily identified.

VII. CONCLUSION

We conclude that critical analysis on applications of outlier detection will help in further research approaches. Outlier information is very useful when data is compared with the original data. The above critical review will help in the further research. Outlier detection methods give a simple & concrete output for the given data. . It has been a great work for those who want to start the research on outlier detection and its domain. The entire work consists different phases and lots of theoretical concepts regarding the Anomalies. In this paper includes the existing outlier techniques and applications where the noisy data exists and also describes critical review on various techniques used in different applications of outlier detection that are to be researched further & they gives a specific type of knowledge based data i.e. more useful in research activities.

VII. REFERENCES

[1]. Anguiulli, F. and Fassetti, F. 2007. Detecting Distance-Based Outliers in Streams of Data.CIKM' 07.Pages 811 - 820.

[2]. S. Ramaswamy, R. Rastogi, and K. Shim.Efficient algorithms for mining outliers from large data sets.pages 427–438, 2000.

[3]. RajendraPamula, Jatindra Kumar Deka, Sukumar Nandi , "An Outlier Detection Method based on Clustering", 2011 Second International Conference on Emerging Applications of Information Technology .

[4]. PrashantChauhan , MadhuShukla , " A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of KMeans Algorithm ", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) , IMS Engineering College, Ghaziabad, India @ 2015 IEEE

[5]. Liu, Hancong, Sirish Shah, and Wei Jiang. "On-line outlier detection and data cleaning." Computers & chemical engineering 28, no. 9 (2004): 1635-1647.

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

[7]. Ghosh, A. K., Schwartzbard, A., and Schatz, M. 1999a. Learning program behavior profiles for intrusion detection. In Proceedings of 1st USENIX Workshop on Intrusion Detection and Network Monitoring.51–62.

[8]. Ghosh, A. K., Wanken, J., and Charron, F. 1998. Detecting anomalous and unknown intrusions against programs.In Proceedings of the 14th Annual Computer Security Applications Conference.IEEE Computer Society, 259.

[9]. SamiranGhosh, SaptarsiGoswami, AmlanChakrabarti, Outlier detection from ETL Execution trace. 2011 IEEE.

[10]. Forrest, S., Esponda, F., and Helman, P. 2004. Aformal framework for positive and negative detection schemes. In IEEE Transactions on Systems, Man and Cybernetics, Part B. IEEE, 357 - 373.

[11]. Gonzalez, F. A. and Dasgupta, D. 2003.Outlierdetection using real-valued negative selection.Genetic Programming and Evolvable Machines 4, 4, 383- 403.

[12]. Lee, W. and Stolfo, S. 1998. Data mining approaches for intrusion detection. In Proceedings of the 7th USENIX Security Symposium. San Antonio, TX.

[13]. Gwadera, R., Atallah, M. J., and Szpankowski, W. 2005b.Reliable detection of episodes in event sequences. Knowledge and Information Systems 7, 4, 415 - 437.

[14]. Chow, C. and Yeung, D. -Y. 2002. Parzen-window network intrusion detectors.In Proceedings of the 16thInternational Conference on Pattern Recognition.Vol. 4.IEEE Computer Society, Washington, DC, USA, 40385.

[15]. Ghosh, S. and Reilly, D. L. 1994.Credit card fraud detection with a neural-network.In Proceedings of the 27th Annual Hawaii International Conference on System Science.Vol. 3. Los Alamitos, CA.

[16]. Brause, R., Langsdorf, T., and Hepp, M. 1999.Neural data mining for credit card fraud detection.In Proceedings of IEEE International Conference on Tools with Artificial Intelligence. 103 - 106.

[17]. Aleskerov, E., Freisleben, B., and Rao, B. 1997.Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelligence for Financial Engineering.220-226.

[18]. V. Ilango, R. Subramanian, V. Vasudevan. A Five Step Procedure for Outlier Analysis in Data Mining. European Journal of Scientific Research ISSN 1450- 216X Vol. 75 No. 3 (2012), pp. 327-339.

[19]. Guttormsson, S., II, R. M., and El Sharkawi, M. 1999.Elliptical novelty grouping for on-line short-turn detection of excited running rotors.IEEE Transactions on Energy Conversion 14, 1 (March).

[20]. Dorronsoro, J. R., Ginel, F., Sanchez, C., and Cruz, C. S. 1997.Neural fraud detection in credit card operations. IEEE Transactions On Neural Networks 8, 4 (July), 827 - 834.

[21]. Bolton, R. and Hand, D. 1999. Unsupervised profiling methods for fraud detection. In Credit Scoring and Credit Control VII

[22]. Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.

[23]. Grzegorz M. Boratyn, Tomasz G, Smolinski, MariofannaMilanova, Jaeek M, Zurada, SudeepaBhattacharyya, and Larry J. Suva.BAYESIAN APPROACH TO ANALYSIS OF PROTEIN PATTERNS FOR IDENTIFICATION OF MYELOMA CANCER.Proceedqs of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2-5 November 2003.

[24]. Zengyou He, XiaofeiXu, Shengchundeng, Discovering Cluster Based Local Outliers.

[25]. Anbarasi. M. S, Ghaayathri. S, Kamaleswari. R, Abirami. I. Outlier Detection for Multidimensional Medical Data. International Journal of Computer Science and Information Technologies, Vol. 2 (1), 2011, 512-516.

[26]. PhyoPhyo San, Sai Ho Ling, Member, IEEE, and Hung T. Nguyen,. Block Based Neural Network for Hypoglycemia Detection. 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011.

[27]. NorizamSulaiman, MohdNasirTaib, SahrimLias, ZunairahHj Murat, SitiArmizaMohdAris, Noor Hayatee Abdul Hamid,.EEG-based Stress Features Using Spectral Centroids Technique and k-Nearest Neighbor Classifier.2011 UKSim 13th International Conference on Modelling and Simulation.

[28]. Chen, D., Shao, X., Hu, B., and Su, Q. 2005.Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. Analytical Sciences 21, 2, 161 - 167.

[29]. Olga Duran and Maria Petrou, A Time-Efficient Method for Anomaly Detection in Hyperspectral Images. IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 45, NO. 12, DECEMBER 2007

[30]. Singh, S. and Markou, M. 2004. An approach to novelty detection applied to the classification of image regions. IEEE Transactions on Knowledge and Data Engineering 16, 4, 396 - 407. To Appear in ACM Computing Surveys, 09 2009

[31]. Tarassenko, L. 1995. Novelty detection for the identification of masses in mammograms.In Proceedings of the 4th IEEE International Conference on Artificial Neural Networks.Vol. 4. Cambridge, UK, 442 – 447

[32]. Diehl, C. and Hampshire, J. 2002. Real-time object classification and novelty detection for collaborative video surveillance.In Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, Honolulu, HI

[33]. Hazel, G. G. 2000. Multivariate Gaussian MRF for multispectral scene segmentation and outlier detection. GeoRS 38, 3 (May), 1199 - 1211.

[34]. Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model.In Proceedings of the IEEE Workshop on Mathematica Methods in Biomedical Image Analysis.IEEE Computer Society, Washington, DC, USA, 3.

[35]. Tao Cheng Zhilin Li, A MULTISCALE APPROACH TO DETECT SPATIAL- TEMPORAL OUTLIERS.
[36]. Bonnie Ghosh-Dastider RAND, Santa Monica OPR, J. L. Schafer, Outlier Detection and Editing Procedures for Continuous Multivariate Data.
[37]. Baker, D., Hofmann, T., McCallum, A., and Yang, Y. 1999.A hierarchical probabilistic model for novelty detection in text.In Proceedings of International Conference on Machine Learning.
[38]. MevlutGullu and Ibrahim Yilmaz, Outlier detection for geodetic nets using ADALINE learning algorithm. Scientific Research and Essays Vol. 5 (5), pp. 440- 447, 4 March, 2010.
[39]. Pokrajac, D., Lazarevic, A., and Latecki, L. J. 2007. Incremental local outlier detection for data streams. In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining(data)
[40]. Janakiram, D., Reddy, V., and Kumar, A. 2006. Outlier detection in wireless sensor networks using Bayesian belief networks. In First International Conference on Communication System Software and Middleware. 1 - 6.
[41]. Branch, J., Szymanski, B., Giannella, C., Wolff, R., and Kargupta, H. 2006.In-network outlier detection in wireless sensor networks.In 26th IEEE International Conference on Distributed Computing Systems.
[42]. Phuong, T. V., Hung, L. X., Cho, S. J., Lee, Y., and Lee, S. 2006. An outlier detection algorithm for detecting attacks in wireless sensor networks. Intelligence and Security Informatics 3975, 735 - 736.
[43]. Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D. 2006.Online outlier detection in sensor data using non-parametric models. In VLDB '06: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 187 - 198.
[44]. Ide, T., Papadimitriou, S., and Vlachos, M. 2007.Computing correlation outlier scores using stochastic nearest neighbors.In Proceedings of International Conference Data Mining. 523 - 528.
[45]. Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys 41.
[46]. Marin, J.M.M., Kerrie, L., and Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions, Vol 25 (Elsevier).