ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Overview of Outlier Detection

Ankita Wagh[1]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Sagar Mhaske[2]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Komal Gaikwad[3]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Trupti Khot[4]
Department of Computer Engineering
JSPM's JSCOE
Pune, India

Abstract: ***In this paper we review the main concept of outlier detection. Outliers were once called as noisy data in statistics, now has turned out to be an important aspect which is being researched in various fields of research and application. Detection of outlier or novelty detection aims in finding patterns in data that deviates from the expected behavior. Arisal of the outliers can be due to mechanical errors, human errors, and instrumental errors or simply due to deviations in the data. Various techniques have been introduced and implemented in order to detect the outliers. It has wide use in variety of applications such as military surveillance for suspicious actions, detection of intrusion in cyber security and fraud detection. Some applications are developed for confidentiality in case of crime and terrorist activities. In this paper we aim to attain better understanding of various methodologies/techniques for outlier detection.***

Keywords: Outlier, Score, Label, NMF, R-NMF, WSN, RNN, PAM.

I. INTRODUCTION

The prime objective of outlier detection is finding the patterns in data that doesn't behave expectedly [11]. The importance of outliers in data lies into the fact that they can be translated into relevant information which in turn can be useful for various applications [11]. In most applications the data is generated through various processes which either could be the observations collected about entities or activities in the system processes. When these processes depict unusual behavior, it leads to creation of outliers [1]. In this paper we are getting to understand the techniques such as distance based technique, one-class vector mechanism, replicator neural networks, cluster analysis, wireless sensor networks [11]. The recognition of unusual characteristics is provided by many applications [7]. Few applications are intrusion detection system, unusual sensor events, medical diagnostics, credit card fraud, law enforcement and so on. In outlier detection the output of an outlier can be of two types:

1. Score:

It is the output which gives the level of outlierness. Tendency of outliers can be determined using score by ranking the data points. This is the most common form of output which provides information about a particular algorithm. But in case of small number of data points, it does not provide precise information.

2. Label :

This kind of output indicates whether a data point is an outlier or not through binary labeling. It may directly return the binary labels or the scores can also be converted into the binary labels. A binary labeling contains less information than a scoring mechanism.

In this paper we hope to analyze different directions of research regarding techniques of outlier detection and applications of these techniques in different areas.

II. LITERATURE SURVEY

Outlier Detection is an important branch in Data Mining. Outlier Detection algorithm have application in several tasks within Data Mining [1]. Data mining is a discovery of Data that provide lot of data from other data pattern. Outlier Detection is referred as Anomaly Detection, Event Detection, Deviant Discovery, Change Point Detection, Fault Detection, Intrusion Detection [11]. A Supervised, Semi-Supervised & Unsupervised Techniques used for Outlier Detection. Outlier Detection is pattern in data that does not confirm nature of normal behavior [1]. The Important aspect of outlier detection Techniques is nature of desired outlier.

Outlier can be classified in following three categories:

» Point Outliers

» Contextual Outliers

» Collective Outliers

Point outlier is Simple type of outlier. It is a focus of majority of research on outlier detection. There is two types of contextual outlier i.e. contextual attribute & Behavioral Attribute. The Contextual attributes are used to determine the context for that instance. The behavioral attribute the non-contextual characteristic. Collective outlier means related data instances are anomalous with entire data set [1].

Outlier Detection techniques:

There are various types of outlier detection techniques. Following are the techniques.

1) Cluster analysis based on outlier detection :

Clustering is the set of cluster which contain all objects in data set. Clustering means grouping set of objects. Cluster analysis is effective sample screened out from original data. This technique is based on cluster analysis. These techniques are used Image Analysis, Pattern Recognition, Bioinformatics and Machine Learning. Cluster analysis is not an algorithm, we can achieve it from various algorithms [4].

Following are the various cluster models:

» Connectivity models : Ex-hierarchical clustering

» Centroid model : Ex-k-means

» Density model : Ex-DBSCAN,OPTICS

» Group model

» Distribution model

2) Distance based technique :

Distance based technique is used to compute the distance. This technique is suitable for situations where observed distribution does not fit in standard distribution [2].

There are two types of distance based technique:

a) k-nearest neighbor :

In case of pattern recognition nearest neighbor algorithm which is non parametric method and that is used in regression and classification. Instance based learnings type is k-nearest neighbor [2].

b) Local outlier factor :

Local outlier factor is proposed by best scientist Markus M. Breunig. It assigns to each of object a degree to be an outlier. And this degree is called LOF of an object. Local outlier factor shares 2 concepts .One is DBSCAN and second is OPTICS. Both are used for local density estimation [2].

3) Replicator neural network :

Replicator neural network is used to provide main outlines of the data records. In RNN neutron stores the pattern based on those pattern they detect the data. The effectiveness of RNN for outlier detection is demonstrated on 2 publically available databases. Best example of RNN is fraud detection like ATM, money banking, mobile phone. RNN is also used to finding outliers [3].

4) Combining Non-negative Matrix Factorization(NMF) with subspace analysis to discover and interpret outliers :

This is the main technique. NMF is a new algorithm where we are using combining approach of SR-NMF and R-NMF to improve the efficiency than existing approaches.

There are the four types of NMF.

» Approximate NMF.

» Convex NMF.

» Non negative rank factorization.

» Different cost function and regularizations.

NMF is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into 2 matrices W and H. With property that all 3 matrices have no negative elements [1].

5) Wireless sensor network:

Wireless sensor network deviate from normal pattern of sensed data is considered as outlier. Wireless sensors network have capability to combine to sense, compute and coordinate their activities with ability to communicate results to outside world. WSN distributed autonomous sensors to monitor physical or environmental conditions such as pressure, temperature and sound [5].

WSN has following application:

» Process management.

» Health care monitoring.

» Forest fire detection.

» Landslide detection.

III. Related Work

Outlier detection has been widely researched and number of approaches has been discovered over the time. Cluster based approach is performed by using Partitioning around Medoids (PAM) and then the small clusters are considered as outlier clusters [4]. After a decade of research, distance based outlier detection algorithms have been proved as scalable and parameter-

free alternative for outlier detection [2]. Replicator Neural Network (RNN) measure outlyingness of data records and is assessed using a ranked score measure [3]. Combining Non- Negative Matrix Factorization (NMF) is an approach used to detect outliers in high dimensional observational space [1]. Another approach is Wireless Sensor Networks (WSN), where sensor reads the normal pattern and deviates it from the data that is considered as an outlier [5].

IV. METHODOLOGY

In this paper we are describing one methodology for discovery of outliers. There are various methodologies for outlier detection like cluster analysis, distance based technique, replicator neural network, and wireless sensor network, combining NMF with subspace analysis to discover and detect outliers. The technique which we are using is NMF. Non negative matrix factorization, it is a group of algorithms in multivariate analysis and linear algebra where X is factorized into two matrices U and V with property that all 3 matrices have no negative elements. This non negativity factor makes resulting matrix easier to inspect. NMF is sensitive to outliers, if fluctuation in data occurs it immediately shows outliers. So we make it robust so that it can show exact outlier or actual outlier. For making outlier robust we are using R-NMF algorithm. R-NMF means robust non negative matrix factorization which is insensitive to outliers.

Algorithm: [R-NMF Algorithm]

Input:

A matrix X of size m x n, m number of features, n number of samples, k the size of the latent space.

Output:

An m x k matrix U and k x n matrix V

R x UV

1: U^0 random m x k matrix

2: i= 1

3: while (no convergence achieved) do

4: $V^i = \text{argmin}_V ||X-U^{i-1}V||_F$

5: $R = X - U^{i-1}V^i$\\R is a residual matrix

6: Let L = {1, 2, . . . , n} be a new ordering of the columns

of R such||R(:, 1)|| (||R(:, 2)|| . . . (||R(:, n)||

7: $X_{-l} = X (:, L \setminus L(1 :l))$

8: $V_{-l} = V (:, L \setminus L(1 : l))$

9: $U^i = \text{argmin}_U|| X_{-l} - UV^i_{-l}||$

10: i= i + 1

Above is the algorithm which we are using to detect outliers and we modify NMF to make them more robust against outliers.

The Algorithm is as follows:

We begin by initializing U in line 1. In line 4 we are solving for V which is minimizing the frobenious norm of $||X-U^{i-1}V||_F$ [1]. In line 5 we are calculating the residual of X and current estimate of $U^{i-1}V$ [1]. In line 6 ranking of the residuals is based on the norm of their column values, L is the index vector of the ranking [1]. In line 7 and 8 we are generating new matrices X_{-l} and

V_{-l} by eliminating first l values of set X and V [1]. We estimate U by minimizing the frobenious norm of X-l and UV^{l}_{-l} in line 9 [1]. We iterate until no convergence criterion is achieved. So in such way detect outlier by using this algorithm.

V. CONCLUSION

Outlier Detection is a significant task in data mining and knowledge and discovery's the size and complexity of data sets increases the need to identify meaningful and genuine outliers will become necessary. The main challenge is to differentiate between genuine and noise outliers. One methodology to distinguish between genuine and noise outliers is to take a multiple subspace viewpoint. A genuine outlier will stand out in multiple subspaces while a noise outlier will be separated from the core data in much fewer subspaces. However the problem in subspace exploration is that current methods are unlikely to scale to high dimensions. The challenging aspect with matrix factorization methods is that they are highly sensitive to outliers. This can be a serious problem whenever there is a mismatch between the data and the proposed model. One way to improve the problem is to use an alternate minimization approach to estimate both the matrix decomposition and the outlier set.

ACKNOWLEDGMENT

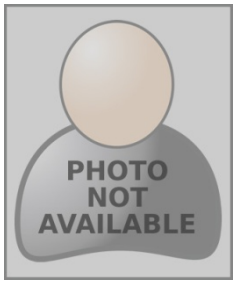
We would like to express our gratitude towards **Prof.H.A.Hingoliwala** whose support and consideration has been an invaluable asset during the course of this project. First and foremost, we would like to thank our guide **Prof.S.M.Shinde** for providing us with their invaluable guidance throughout the course of this project. It would have been an almost impossible task to complete this project without his support, motivation, valuable suggestions and criticism. We convey our gratitude to our respected **Head of Department, Prof.H.A.Hingoliwala for** his motivation, guidance, and criticism and also for providing various facilities, which helped us greatly in the course of this project.

And, last but not the least we would like to thank **Principal M.G.Jadhav** and all the teaching, non-teaching staff of Computer Department and also our friends for directly or indirectly helping us for the project completion and all the resources provided.

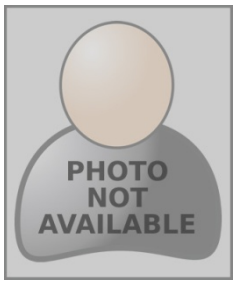
References

1. Fein Wang, Sanjay Chawla, Didi Surian, "Latent Outlier Detection and the Low Precision Problem", University Of Sydney & NICTA Sydney, Australia.
2. Gastavo H. orair, Carlos H. C. Teiexeira, Wages Meira Jr. ,Ye Wang, Srinivasan Parthasaratng,"Distance-Based Outlier Detection Consolidation & Renewed Bearing".
3. Simon Hawkins, Hongxing He, Graham Williams and Rohan Baxter, "Outlier Detection Using Replicator Neural Networks" ,CSIRO Mathematical & Information Sciences.
4. Vijay kumar, Sunil Kumar, Ajay kumar Singh "Outlier Detection:A Clustering-Based Approach" , International Journal Of Science and Moldering Engineering(IJISME) ASS N:2319-6386,,Volume-1,Issue-7,June 2013.
5. Gaurav Sahni,Sonia Sharma,"Study of Various Anomalies & Anomaly Detection Methodologies in Wireless Sensor Network",CSE Dept. , Kurukshetra University India.
6. M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In IEEE International Conference on Data Mining (ICDM), 2000.
7. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Comput. Surv., 41(3), 2009.
8. S. Chawla and A. Gionis. K-means: A unified approach to clustering and outlier detection. In SIAM International Conference on Data Mining (SDM SIAM), 2013.
9. T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. Knowledge and Information Systems (KAIS), 32(1):25–52, 2012.
10. Frank and A. Asuncion. UCI machine learning repository, 2010.
11. Outlier Detection: Principles, Techniques and Applications Sanjay Chawla and Pei Sun School of Information Technologies University of Sydney NSW, Australia chawla|psun2712@it.usyd.edu.au

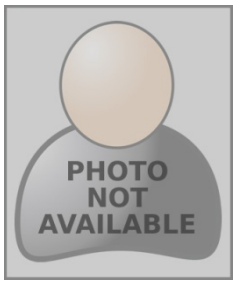
AUTHOR(S) PROFILE



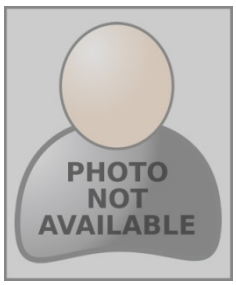
Ms. Ankita Wagh, currently BE student in the Computer Science from Jaywantrao Sawant College of Engineering. And my research interested areas are in the field of Java, Data mining.



Mr. Sagar Mhaske, currently BE student in the Computer Science from Jaywantrao Sawant College of Engineering. And my research interested areas are in the field of Java , Data mining and web development.



Ms. Komal Gaikwad, currently BE student in the Computer Science from Jaywantrao Sawant College of Engineering. And my research interested areas are in the field of java , Network Security, and Data mining.



Ms. Trupti Khot, currently BE student in the Computer Science from Jaywantrao Sawant College of Engineering. And my research interested areas are in the field of java , Network Security, and Data mining