A Review on Cluster Creation for High Dimensional Discrete Data And Pattern Based Anomalous Topic Discovery

Ms. Urwashi Patil[1], Prof.M.B.Vaidya[2]

[1] PG student, Department of Computer, AVCOE, Maharashtra, India
[2]Assistant Professor, Department of Computer, AVCOE, Maharashtra, India

ABSTRACT

Generally, discovering of an abnormal data i.e. anomalies from discrete data leads towards the better understanding of atypical behavior of patterns and to identify the root of anomalies. Anomalies can be defined as the patterns that do not have normal behavior. It is also called as outlier detection. Anomaly detection techniques are mainly used for fraud detection in credit cards, bank fraud, network intrusion [15] etc. It can be referred as, novelties, deviation, exceptions or outlier. Such type of patterns cannot be observed to the analytical definition of an outlier, as unusual object till it has been integrated properly. A cluster analysis method is used to detect micro clusters formed by these anomalies. There are various methods existed for detecting anomalies from datasets which only detects the individual anomalies. Problem with individual anomaly detection technique that detects anomalies using the entire features typically fail to detect such anomalies. A method to detect cluster of anomalous data combine manifest atypical section of a small subset of features. This method uses a null model to for typical topic and then separate test to detect all clusters of abnormal patterns.

Keywords:-*Anomaly Detection, Pattern Detection, Topic Models, Topic Discovery*

1. INTRODUCTION

Particularly, in data analysis anomalies like, outlier, deviation, exceptions etc are important concepts. Data objects to be considered as outlier if it has some fluctuation from the regular data behavior in specific region. It means that the data object from the given dataset has “dissimilar” behavior. To detect such type of objects from the given dataset is a very important and crucial task as they need to treat differently from the other data. Anomaly detection is widely used in credit card fraud detection [14], bank fraud detection [11], Whole-genome DNA matching, filtering of ECG signals. AD is the problem has become recognized rapidly developing topic of the data analysis. Our main purpose is to report specific features of widely known analytical and machine learning method used to detect anomalies. The goal is to detect anomalies form the dataset which consists of some normal and some abnormal instances. Sometimes it happens that there is no idea about normal instances which tends to make critical task for identifying abnormal instances from the given dataset. In computer network [15], anomalous patterns traffic could be mean as hacked computer is sending sensitive information to the unauthorized destination [16].

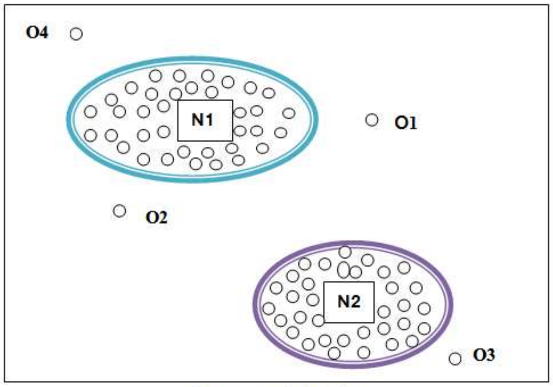

Fig -1 Anomaly Detection

Figure 1 shows anomalies in a 2-dimension.It is two dimensional plane of data sets. N1 and N2 are two normal regions. According to the observations most of data sets lie in these regions. If we observe carefully then we came to know that point's o1 and o2, o3, o4 are the points which not lies in normal regions. They are far away from the normal regions. So we can say that they are anomalies. Figure 1 represents the very simple example of outliers in 2-D plane. Anomalies may be introduced in the data for so many reasons and they are not noise which must be eliminated. Anomalies might be evoked in the data for so many reasons, such as malevolent activity, e.g., credit card fraud, terrorist activity, intrusion or breakdown of a system[14]. But the communal component of all is that they are fascinating to the expert. The interestingness of it or its real life relevancy of outliers is a feature film of outlier detection [13]. The main aim of AD is to find out patterns in data sets that shows unexpected behavior. It owns all-encompassing usage in a huge variety of applications. This researched problem has immense use in a wide variety of application domains such as credit card[14], insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas. In computer data irregular traffic pattern may be shows that a computer is hacked. It is sending out highly sensitive data. An anomalous MRI picture may shows presence of cancerous tumors. Outliers in transactions related to credit card data could identity theft and so on. Mainly, Anomaly detection is related to but distinct from noise removal. Novelty detection is related to the anomaly detection which detects the previously unobserved patterns in the data. Detecting anomalies is the technique for detecting individual sample anomalies. In data mining, fraud detection is nothing but the classification of data. Previously, Mixture of Gaussian Mixture Models is utilized for group anomaly detection [2]. This technique assumes each data point belongs to one group and the all points in the group are modeled by MGMM. Futhermore, idea of MGMM is extended to FGM i.e. Flexible Genre Model. it treats the mixing proportions as random variables considered as normal genres. There are some limitation for MGMM and FGM is that only working on high dimentional feature space. Therfore, it may be inaccurate when anomalous patterns lies on low dimentional feature subspaces. Another method introduced in[3], implemented to overcome the limitations of previous techniques. This is network analysis method[16] to detect similar nodes for computing anomaly scores for hidden groups[15]. Prevous methods for anomaly detection does not have an algorithmic procedure for discovering “hard” anomaly clusters individualy[4]. this methods only detects the individual anomalies. In[1], there exist a method for detecting cluster or a group of an anomalies. This method can helps to detect abnormal behaviour of patterns as well as to identify the root or sources of anomalies. This propose method considered sufficiently characterised normal data. it uses a null model in training phase to detect possible clusters of anomalous patterns in different test batch. This frameork has important applications in various domain for example in, scientific or business related applications. Identification of anomaly clusters have many applications to detect similar pattterns in malware and spyware to dignose the sources of attacks,studying patterns of an anomalies to discover the customer behaviour.

2. RELATED WORK

In this section we are going to discussed related work about existing techniques for anomaly detection.
They are explain as following:

A. Outliers or Anomaly Detection

Anomaly or outlier pattern are those which depicts the abnormal task than the other patterns of same dataset. the above figure depicts dataset which having two i.e. N1 and N2 regions. From the observation on both regions it seems that O1,O2,O3 and O4 are the points far away from the regions. Hence, those points are called as anomalies in dataset. anomalies discover in the data for variety of reasons. It can be a malicious activity such as, credit card frauds, cyber intrusion, some terrorist activity etc. AD is distinct from the noise removal as well as noise accomendation as both are deal with unnecessary noisy data. Novelty detection is way of detecting emergent and novel patterns in the data. The difference between anomalies and the novel pattern detection is that novel pattern is characterised into normal model when it is detected. There certain limitations in detection of anomalies such as, it is complicated to define normal behaviour of patterns or to define noemal region. Binding of every possible normal behaviour is impossible. Also variations of malicious attackers to make anomaly observations like a normal when they result from malicious actions. Noise in the data tends to be similar to the original anomaly therefore it is difficult to distinguish and remove.

B. Group Anomaly Detection

B.1 MGMM

MGMM is Mixture of gaussian Mixture Model used for group anomaly detection in[2]. In this technique assumes each data point related to one group and all the points in that groups are modeled by group's gaussion mixture model. MGMM model is effective for uni-modal group behaviours. It is extended as GLDA i.e. Gaussian LDA to handle mlti-modal group behaviour. Both techniques detects point-level and group level anomalous behaviour.

B.2 FGM

Another technique is Flexible Genre Model. FGM treats mixing proportion as random variables. Random variables are modified on possible normal genres. This method assumes the membership of each data point which is known as, apriori[3]. Practically it is hard to clustering data into groups of preceeding to applying FGM as well as MGMM mechanism.

C.GLAD:Group Anomaly Detection in social Media Analysis

Author R.Yu, X.He, Y. Liu proposed the problem of group anomaly detection in social media analysis. To define group anomaly they were identified the group membership as well as the role of individual. GLAD model is also called as Bayes model used for detecting group anomaly. It utilises both pair-wise and point-wise data to automatically guess the membership of group as well as role of individuals. Extension for GLAD model is d-GLAD model ustilised to handle sampling time series. For the smapling of time series variational bayesian and Monto Carlo sampling model is used. Synthetic datasets as well as real world social media datasets are used to evaluate the performance of GLAD and d-GLAD model. GLAD model successfully detects the anomalous papers from scientific publication dataset with included anomalies whereas, d-GLAD extracts the official relationships changes in the councelling related to the political events[10].

In[4],OCSMM i.e one-class support measure machine algorithm used to detect anomalies in group. It handles the aggregate behaviour of data points. Distribution of groups are represented using RKHS through kernel mean embeddings. Author K. Muandet and B. Scholkopf extended the relationship between OCSVM and the KDE to the OCSMM in the relation of variable kernel density estimation, overcoming the gap between large-margin approach and kernel density estimation.

D.Ruled Based Anomalous Pattern Discovery

A rule-based anomaly pattern discovery is discussed in[15],to detect anomalous patterns rather than the pre-defined anomalies. In this anomalous pattern discovery each pattern is summerised by a rule. In implementation phase it consist of one or two components. In this mechanism of ruled based anomalous pattern discovery, rule is simply set of possible values which subset of categorical features[9]. This approach required to wary certain risks of rule-based anomaly pattern detection. Hence there have to find anomalous patterns rather than isolated anomalies. To monitor healthcare data to check irregularities disease outbreak detection system is discussed in[5]. In[5] research paper, baseline method is replaced with Bayesian network[15]. Baysian network generates baseline distribution by taking the joint distribution of data. The WSARE algorithm can detects the otbreaks in simulated data with earlier possible detection. Detecting anomaly pattern in Categorical Datasets is represented in [6].

E. Clustering with MapReduced Strategy

N.Gosavi,et al. [17], proposed a protocol to solve privacy of database confidentiality which is affected while transforming database from one to another. Proposed protocol is generalised k-anonymous and confidential databases. Several techniques hs been discussed by them such as, randomization, k- anonymity etc. In randomisation a way of protecting the user from learning sensitive data is given. It is simple technique because it does not require to knowledge of other records. They defined applications of their proposed work in military application or health care system. But there are some limitations identified with this approach is not sufficient protocol as if a tuple fails to check, it does not insert to the database and wait until k- 1.because of this much of long process waiting time also gets increase. Some imporant problems are planned in their futur work, invalid entries database implemenation, to improve efficiency of protocol in terms of number of messages exchanged etc.

Y.Patil, M. B. Vaidya [18], discussed about K-Means Clustering Algorithm over a distributed network. They have utilised map-reduce technique for proposed system implementation. Proposed algorithm robust and efficient system for grouping of data with same characteristics but also reduces the implementation costs of processing such huge volumes of data. They predicted that, for text or web documents K-means clustering using MapReduce is can be more suitable. Their main focused is over a distributed environment using Apache Hadoop. In future work clustering with hadoop platform is suggested by them.

4. CONCLUSIONS

In this review paper we have discussed some existing technique used for outlier detection[13], novelty detection and anomaly detection etc. In this survey we found that anomalies are the patterns which have abnormal behaviour than the regular patterns. Previous methods used in anomaly detection have certain limitation as, only individual anomaly can be detected, some approaches like, MGMM and FGM can efficiently works on high density dataset. There are some techniques such as GLAD, d-GLAD, OCSMM which discovers the behaviour of anomalies in group. WSARE algorithm used in rule based anomaly pattern discovery. It detects the anomaly in categorical dataset. According to our analysis from this literature review we plan to design a system that can efficiently works on synthetic as well as real datasets which can be capable of identifying group/cluster of anomalies with low density.

5. ACKNOWLEDGEMENT

A very firstly I gladly thanks to my project guide Prof.M.B.Vaidya, for his valuable guidance for implementation of proposed system. I will forever remain a thankful for their excellent as well as polite guidance for preparation of this report. Also I would sincerely like to thank to HOD of computer department Prof.R.L.Paikrao and other staff for their helpful coordination and support in project work.

6. REFERENCES

[1] Hossein Soleimani, and David J. Miller ,"ATD: Anomalous Topic Discovery in High Dimensional Discrete Data",IEEE transaction on knowledge and data engineering,2016

[2] L. Xiong, s. P. Barnaba, J. G. Schneider, A. Connolly, and V. Jake, ´ "Hierarchical probabilistic models for group anomaly detection," in International Conference on Artificial Intelligence and Statistics, pp. 789–797, 2011.

[3] L. Xiong, B. Poczos, and J. Schneider, "Group anomaly detection ´ using flexible genre models," in Advances in neural information processing systems, pp. 1071–1079, 2011.

[4] R. Yu, X. He, and Y. Liu, "GLAD : Group Anomaly Detection in Social Media Analysis," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 372–381, 2014.

[5] K. Muandet and B. Scholkopf, "One-class support measure ma- ¨ chines for group anomaly detection," in 29th Conference on Uncertainty in Artificial Intelligence, 2013.

[6] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Rule-based anomaly pattern detection for detecting disease outbreaks," 2002.

[7] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," 2003.

[8] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," 2008

[9] E. McFowland, S. Speakman, and D. Neill, "Fast generalized subset scan for anomalous pattern detection," Journal of Machine Learning Research, vol. 14, no. 1, pp. 1533–1561, 2013.

[10] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," 1998.

[11] X. Dai, Q. Chen, X. Wang, and J. Xu, "Online topic detection and tracking of financial news based on hierarchical clustering," in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, pp. 3341–3346, 2010.

[12] Q. He, K. Chang, E.-P. Lim, and A. Banerjee, "Keep it simple with time: A reexamination of probabilistic topic detection models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 10, pp. 1795–1808, 2010.

[13] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, 2004.

[14] A.Srivastava and A. Kundu, "Credit card fraud detection using hidden Markov model," IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008

[15] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection," in Recent Advances in Intrusion Detection, pp. 203– 222, 2004.

[16] F. Kocak, D. Miller, and G. Kesidis, “Detecting anomalous latent classes in a batch of network traffic flows,” in Information Sciences and Systems (CISS), 2014 48th Annual Conference on, pp. 1–6, 2014.

[17] N.Gosavi, S.H.Patil, “Generalization Based Approach to Confidential Database Updates,” in International Journal of Engineering Research and Applications (IJERA), vol.2, Issue 3, pp.1596-1602,May-June 2012.

[18] Y.S.Patil, M.B.Vaidya, “K-means Clustering with MapReduce Technique,” in International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol.4, Issue 11, November 2015.