A Discovery of Group Anomaly and Emerging Topics Using ATD Approach

[1]Shewale Yogita, [2]Prof. Shinde Jayashri

[1,2]Department Master of Computer Engineering, Late G. N. Sapkal College of Engineering Nashik, Maharashtra, India.

[1]*yogitashewalenet.shewale@gmail.com* , [2]*jv.shinde@rediffmail.com*

Abstract **In the area of diverse research, anomaly detection is an important problem. Anomaly is the pattern that does not conform to the expected behavior. It can refer as outlier, exceptions, surprise etc. Anomalies can be translated to real time entity such as, fraud detection, cyber intrusion etc. Many types of anomaly detection techniques have been proposed in literature but that only capable of detecting individual anomalies. In this paper we proposed ATD algorithm to detect cluster of anomalies. Individual anomaly detection technique fails to detect atypical pattern that exhibit on salient subset of vary high dimensional feature space. Our proposed algorithm consists of two steps. First is the training step in which we learn PTM as our null model M_0 to generate all document in test set. Second is the detection phase in which we utilized document bootstrapping algorithm for clustering of candidate documents (S) in the test set. Furthermore, as a part of contribution we focus on emergence of topics signaled by social aspects by discovering links between social users. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in documents. For experimental result analysis we used NASA and BBC dataset. With experimental results we aim to represent that the proposed approach can efficiently detects a cluster of anomalies and emerging topic in test set.**

Keywords: Anomaly Detection, Pattern Detection, Topic Models, Topic Discovery.

I. INTRODUCTION

AD techniques typically detect individual sample anomalies. In this work, however, we focus on detecting abnormal patterns exhibited by anomalous groups (clusters) of samples. An anomalous cluster is a set of data samples which manifest similar patterns of a typicality. Each of the samples in such a cluster may not be highly atypical by itself, but, when considered collectively, the cluster demonstrates a distinct pattern which is significantly different from expected (normal) behavior. In this paper, we propose a framework to detect such groups of anomalies and the atypical patterns they exhibit. Moreover, we consider the case where the anomalous pattern may manifest on only a small subset of the features, not on the entire feature space; i.e., samples in the anomalous cluster may be far apart from each other measured on the full feature space, but on a subset of the feature space (the salient features), they exhibit a similar pattern of abnormality. In addition to detecting atypical clusters, our proposed method identifies each cluster’s salient feature subset.

In some cases, no prior knowledge about normal behavior is available, and the goal is to detect anomalies (outliers) in a single data set consisting of normal and possibly abnormal instances, without any annotation of which samples are normal. More typically, and as we assume here, there is a collection of normal data which sufficiently characterizes normal behavior. In the training phase, we use this data to build a (null) model. Then, in the detection phase, this model is used as a reference to help detect (possible) clusters of anomalous patterns in a different (test batch) data set. Our proposed framework has significant applications in a variety of domains. For instance, consider an open repository of scientific or business related articles. A company may try to post articles on this repository to promote its products or services. However, to avoid being easily detected by normal advertisement blocker services, the articles are written in such a way that they match the normal articles on that repository in form and content. Only a small part of these advertising articles promote the company's services. In this case, we can identify that company's infiltration by detecting clusters of such articles.

In order to do so, we first use a sub-collection of normal articles from that repository as our training set to learn the normal topics (null model). Then, using that null model, our algorithm detects clusters of such advertising articles within the full repository, the anomalous topic of each cluster (the product or service they promote), and the keywords representing that topic.

Some other potentially important applications of our framework are: detecting similar patterns in malware and spyware (that were uploaded to a public software tool repository) to identify sources of attacks; studying patterns of anomalies in consumer behavior to discover emerging consumer trends; finding shared patterns of tax avoidance to reveal loopholes in the law; and detecting organized malicious activities in social media.

II. LITRATURE SURVEY

Srivastava, A. Kundu, S. Sural et al. [1], proposed an application of HMM in credit card fraud detection. HMM is assumptive process of representation of different steps of credit card transaction processing. In this paper, an HMM is initially trained with the normal behavior of a cardholder. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. A simulator is used to generate a mix of genuine and fraudulent transactions. The number of fraudulent transactions in a given length of mixed transactions is normally distributed with a user specified mean and standard deviation, taking cardholder's spending behavior into account. In a typical scenario, an issuing bank, and hence, its FDS receives a large number of genuine transactions sparingly intermixed with fraudulent transactions. The genuine transactions are generated according to the cardholders' profiles. The cardholders are classified into three categories as mentioned before the low, medium, and hs groups.

W.K. Wong, A. Moore et al [3], proposed a rule based anomaly detection algorithm. It is used to characterize each anomalous pattern with a rule. The proposed algorithm is compared against a standard detection algorithm by measuring the number of false positives and the timeliness of detection. This algorithm considers all patient records falling on the current day under evaluation to be recent events. In this paper to detect anomaly, there is a need of such concept that something being normal. Also there is requirement of account for environmental factors as weekend versus weekday differences in the number of cases. Other WSARE is abbreviated form for "What's strange about recent events". It operates on discrete data sets with the aim of finding rules that characterize significant patterns of anomalies. Due to computational issues, the number of components for these rules is two or less.

K. Das, J. Schneider et al. [4], suggested a method for detecting patterns of anomalies in categorical datasets. Local anomaly detector identifies individual records with anomalous attribute values and then detect pattern where the number of anomalous record is higher than expected. In this paper, author represented that the proposed methodology able to accurately detect anomalous patterns in real world hospital container shipping and network intrusion data. The proposed APD is orthogonal to the local anomaly detection method.

J.Allan, R. Papka etc. [5], discussed about the problems of new event detection and event tracking within a stream of broadcast news stories. Before, looking at subsequent stories, decision about one story is taken already. They represented a method based on miss and false alarm rates. In this boostrap method is used to produce performance distribution for topic detection. The TDT tasks and evaluation were developed by a joint effort between DARPA. The group involved in the tasks found that the "State of the art" is capable of providing adequate performance for detection and tracking events, but there is high enough failure rate to warrant significant research into how algorithms can be advanced.

X. Ying, Q. Cai Chen [6], proposed TDT approach to the vertical search engine in financial field. Results are grouped into multiple topics with stock as unit. In this paper, author proposed clustering method called as, hierarchical agglomerative. They also proposed online topic detection and topic tracking with the proposed approach. For splitting, agglomerative hierarchical clustering method into two steps and considers the time factor. In final study the effect on the similarity between two stories or topics. The proposed method limited only for tracking and detection of financial news.

A.P. Dempster, D.B.Rubin [10], represents the general approach to iterative computation of maximum-likelihood estimates when the observation views incomplete data.

III. SYSTEM ARCHITECTURE

Our anomalous cluster detection approach consists of two fundamental steps repeatedly applied to the test batch:
i) determining the best current candidate anomalous cluster;
ii) determining whether this candidate cluster is anomalous.
Note that we do not presume that any anomalous clusters actually exist in the test data. In this paper, we propose statistical tests to accomplish both these steps; i.e., to determine which samples significantly belong to the best current cluster candidate and to test whether the candidate exhibits a statistically significant degree of a typicality relative to the null model.

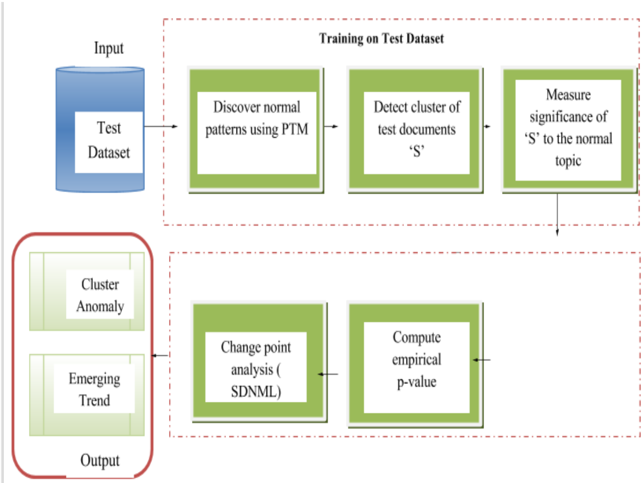
Input
Training on Test Dataset
Test Dataset
Discover normal patterns using PTM
Detect cluster of test documents 'S'
Measure significance of 'S' to the normal topic
Cluster Anomaly
Emerging Trend
Change point analysis (SDNML)
Compute empirical p-value
Output

Fig. 1 System Block diagram

We choose PTM over LDA as the topic model for our ATD algorithm for a number of reasons. First, because PTM typically achieves better generalization accuracy than LDA and it automatically estimates the number of normal topics, unlike LDA, which requires this number to be set by a use. Note that model order selection is a crucial step in anomalous topic discovery. Specifically, since significance of any anomalous topic will be measured with respect to the null model (normal topics), either under or overfitting the null can lead to false discovery of anomalous clusters due, respectively, to limited modeling power or to poor generalization. Moreover, PTM, unlike LDA, identifies a highly sparse set of topic-specific (salient) words for each topic.

This makes PTM a natural fit for our ATD algorithm as we assume that the anomalous topics manifest on a very low dimensional subspace of the full word space. PTM, with its sparse topic representation, is expected to have an inherent performance advantage over LDA, which uses all the words in the dictionary to define topics. In fact, this is supported by our experimental results in the sequel.

Our anomalous topic discovery algorithm consists of two main parts: First, in the training step, we learn PTM as our null modelM0, with M its estimated number of topics. The null hypothesis is that all documents in the test set were generated by the null model. Second, in the detection phase, under the alternative hypothesis, we posit that a cluster of documents in the test set may contain an additional topic.

IV. CONCLUSION

In this paper we proposed ATD algorithm to detect cluster of anomalies. Individual anomaly detection technique fails to detect atypical pattern that exhibit on salient subset of vary high dimensional feature space. Our proposed algorithm consists of two steps. First is the training step in which we learn PTM as our null model M_0 to generate all document in test set. Second is the detection phase in which we utilized document bootstrapping algorithm for clustering of candidate documents (S) in the test set. Furthermore, as a part of contribution we focus on emergence of topics signaled by social aspects by discovering links between social users. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in documents. For experimental result analysis we used NASA and BBC dataset. With experimental results we aim to represent that the proposed approach can efficiently detects a cluster of anomalies and emerging topic in test set.

V. ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on “Anomalous Topic Discovery (ATD)”.
I would like to take this opportunity to thank my internal guide Prof. J. V. Shinde for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful.

REFERENCES

1). 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.

2). H. Soleimani and D. J. Miller, “Parsimonious Topic Models with Salient Word Discovery,” Knowledge and Data Engineering, IEEE Transaction on, vol. 27, pp. 824–837, 2015

3). W. Wong, A. Moore, G. Cooper, and M. Wagner, “Rule-based anomaly pattern detection for detecting disease outbreaks,” 2002.

4). K. Das, J. Schneider, and D. B. Neill, “Anomaly pattern detection in categorical datasets,” 2008

5). 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.

6). 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.

7). M. Zhao and V. Saligrama, “Anomaly Detection with Score functions based on Nearest Neighbor Graphs,” in Advances in neural information processing systems, pp. 2250–2258, 2009.

8). L. M. Manevitz and M. Yousef, “One-Class SVMs for Document Classification,” Journal of Machine Learning Research, vol. 2, pp. 139– 154, 2001.

9). 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.

10). A.P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society., vol. 39, no. 1, pp. 1–38, 1977.

11) X.-L. Meng and D. Van Dyk, “The EM algorithm–an old folk-song sung to a fast new tune,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 59, no. 3, pp. 511–567, 1997.

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) H. Soleimani, D.j. Miller, “ATD: Anomalous Topic Discovery in High Dimensional Discrete Data”, IEEE transaction on knowledge and data mining.