Cluster Creation for High Dimensional Discrete Data and Pattern Based Anomalous Topic Discovery

U.V.Patil[1], Prof.M.B.Vaidya[2]
[1, 2] Department of Computer Engineering
[1, 2] AVCOE, Sangamner, Savitribai Phule Pune University,Pune India.

Abstract- *Anomaly detection is the technique in which individual anomalies are discovered. In proposed system, group or clusters of anomalies are detected. Abnormal patterns get saved into separate cluster. An anomalous cluster is the set of data points which manifests the similar pattern of a typicality. PTM model is used to detect normal topics from given dataset. Some existing approaches work against to detect individual anomalies. There are some limitations for existing system as they detect only individual anomaly and it can efficiently work on high density dataset. Proposed method can detect anomalies by discovering salient feature subsets and detecting clusters of anomalies. This system contribute hadoop platform for implementation.*

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

I. INTRODUCTION

In this work, proposed system aim to detect anomalous topic from batch of text documents. Proposed system design algorithm based on topic models. Previously, there are several methods have been introduced for detection of an anomalies but there limitation is they can detect only a single anomaly which is more time efficient task. In our proposed approach PTM model is introduced. It controls the number of free parameters in the model by balancing complexity and goodness of dataset. Besides that, PTM proposes only sparse subset of topics present in each document. We have utilized PTM due to its better generalization accuracy than LDA model. LDA is also a parameter-rich model, when it is applied to high dimensional problems such as, text documents which may result into poor generalization and semantically unpredictable topics [16]. PTM can automatically estimate the number of normal topics. In anomalous topic discovery, model order selection is challenging task. To compute implication of any anomaly topic will be measure with the null model, either overfitting the null can lead to false discovery of anomalous clusters due, respectively, to limited modeling power or to poor generalization. The proposed ATD algorithm selects documents which contains normal documents. Our main is to discover any or all patterns in the test corpus which are anomalous with respect to the normal topics. In each step, `S' is the candidate anomalous cluster which outperforms maximum "deviations" than normal topics. Under the null and alternative models, a candidate document is more relatively changes. d* is the candidate document which is required for cluster significance computation. D* is belonging to 'S'. After cluster exceeds its capacity, author determines whether the anomalous topic exhibited by the documents in that cluster is significant. An algorithm similar to the procedure for generating bootstrap documents they described a procedure to generate |S| bootstrap documents (Sb) based on the null distribution from a collection of validation documents and compare the likelihood ratio score of this bootstrap cluster with that of the candidate cluster[1].

In experimental setup phase, performance of algorithm compared with baseline methods on synthetic data set and two text corpora dataset. Synthetic dataset is generated based on 10 normal topics on a dictionary with 3000 unique words. As overall discussion, all data sets, that ATD can accurately detect the anomalous topics and their salient features. In this paper proposed statistical test can also determine significance of each detected cluster efficiently and with low false detection rate.

II. RELATED WORK

Anomaly pattern are those which depicts the abnormal task than the other patterns of same dataset. The above figure 1 depicts dataset which having two regions such as, N1 and N2. 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. For various reasons anomalies are discoverd from the data. 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 adaption 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. The task of anomaly detection is complicated due to normal behaviour of patterns or normal regions are defined in

it. Binding of every possible normal behaviour is impossible. Also variations of malicious attackers to make anomaly observations as 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. MGMM is Mixture of gaussian Mixture Model used for group anomaly detection in[2]. 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. 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.

The problem of group anomaly detection in social media analysis is discussed in [4]. To define group anomaly they were identified the group membership as well as the role of individual. GLAD model is used to detect group anomaly which is also known as Bayes model. It utilises both pair-wise and point-wise data to automatically guess the membership of group as well as role of individuals. d-GLAD model is extension to the GLAD model which is ustilised to maintain sampling time series. To monitor healthcare data to check irregularities disease outbreak detection system is discussed in[5][6].

In this mechanism of ruled based anomalous pattern discovery, rule is simply set of possible values which subset of categorical features[7]. This approach required to wary certain risks of rule-based anomaly pattern detection. Hence there have to find anomalous patterns rather than isolated anomalies.In ruled based anomalous pattern discovery, rule is the simple set of possible values which subset of categorical features[8]. This approach required to vary certain risks of rule-based anomaly pattern detection. Hence, there have to find anomalous patterns rather than isolated anomalies[9].
For the sampling of time series variational bayesian and Monto Carlo sampling model is used. To evaluate the performance of GLAD and d-GLAD, synthetic as well as real world social media datasets are used. 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].

OCSMM is One-Class Support Measure Machine algorithm used for group anomaly detection. It maintains aggregate behaviour of data points of anomaly dataset. Distribution of groups are represented using RKHS through kernel mean embeddings.

To detect anomalous patterns rather than the pre-defined anomalies a rule-based anomaly pattern discovery is discussed in [12]. In anomalous pattern discovery each pattern is summerised by a rule. There are two components included in implementation phase. A baseline method is replaced with Bayesian network [13]. It generates baseline distribution by taking the joint distribution of data. The WSARE algorithm detects the breaks in simulated data with earlier possible detection. Detecting anomaly pattern in Categorical Datasets is represented in [14].

III. PROBLEM DEFINITION

From analysis of existing anomaly detection techniques, several methods have been proposed to detect anamalies from dataset which detects individual anomalies. Hence, to proposed a technique to detect anomalies cluster also to enhance the performance of system utilise hadoop platform.

IV. SYSTEM ARCHITECTURE

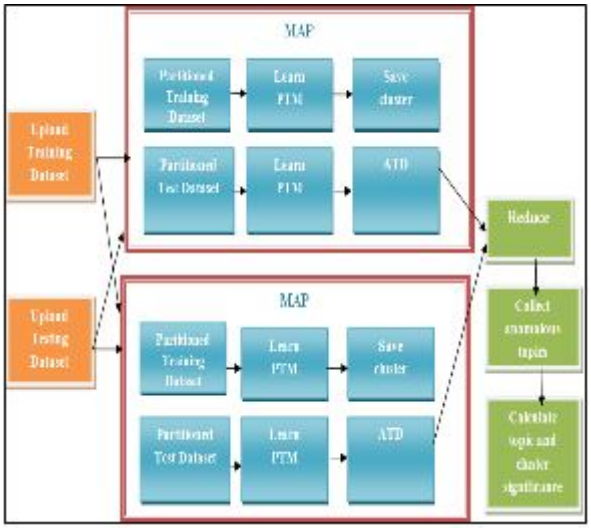

Figure 1. System Architecture

1. Dataset Pre-processing:

In this phase, user uploads the dataset which get pre-processed by system. Input dataset contains various 20 categories news which get sort as per category package into list after pre-processing of dataset.

After pre-processing of data, different categories from generated list are selected for training and testing. Categories selected for training and testing are totally different from each others.

2. Training phase:

Training phase is carried out from following steps:

Phase extraction:

In this phase, important phrases get extracted by applying stemmer and stopword algorithms. In stemmer algorithms, variant forms of a word are reduced to a common form where, in stopword algorithm words such as, the, is, at, which, and on which can cause problems when searching for phrases that include them are removed.

TF-IDF(Term frequency identification and inverse term frequency identification):

In this phase, extracted phrases and words are taken as an input and frequency of each word is evaluated. In term frequency identification, we wish to determine which document is most relevant to the query and simply avoiding documents that do not contain query terms in it.

To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency.

In inverse term frequency identification, inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idf\ (t,D) - log \frac{N}{|\{d \in D : t \in d\}|}$$

Where,
N: total number of documents in the corpus N= |D|
|{d D : t d}| number of documents where the term `t' appears. If the term is not in the corpus, this will lead to a division-by-zero.

To test the performance of proposed system we have used 20-Newsgroup dataset. It contains 2-different categories.

Relevance score calculation:

It gives the relevant document order such as, result that is most relevant to TF-IDF in the search is the first item in the search return sequence, and the least relevant is the last.

3. Testing Phase:

PTM analysis:

In this phase, PTM i.e. the concept of Parsimonious Topic Models is utilised. It controls the number of free parameters in the model by balancing model complexity and goodness of fit to the data set used for learning the model.

Using PTM normal topics from input dataset are extracted. It proposes only sparse subsets of topics are present in each document, with the rest of the topics having zero proportion. PTM achieves better generalization accuracy (classification and test set loglikelihood) than LDA evaluated on multiple text corpora. Objective function of PTM optimized with respect to the number of topics present in the dataset.

Term frequency identification:

Similar task like training phase takes place in testing phase i.e. extraction of frequent words.

Relevance score calculation:

In this phase, document relevance score is calculated which is required for document cluster creation.

Cluster creation:

In this phase, relevance score of test documents is matched with the relevance score of training documents. If there any matches found then similar documents are bind into a cluster else a new cluster is generated for every unmatched document.

If cluster limit is over then new cluster is created.

Anomaly detection:

After successful document cluster creation, further aim is to identify anomaly from them.

The objective is to detect the clusters of anomalous samples in the test batch and identify the salient feature subset for each such cluster. From the set of all clusters, data point

which exhibits the pattern with maximum "deviance" from normal topics.

Then, we conduct a statistical test to measure the significance of S and the topic exhibited by it, compared to the normal topics hypothesis. If the cluster candidate is determined to be significantly anomalous, we declare it as detected.

We remove all documents in S from the test set, and then repeat this process until no statistically significant anomalous topic is found.

4. Hadooop platform:

We have developed this system on hadoop platform. The system efficiency can be compared with respect to the execution time required on hadoop platform and without hadoop platform. Along with the execution time memory analysis is also compared.

V. ALGORITHMS

Notations:
S is the test doc. Set= Dt
M: Normal topics
J: indexing for normal topics such as {1,2,……M}
N: unique words in the dictionary
Ld: discriminate factor
Document d consist of Ld{wid…… wLdd }
M0{null model} including some constant and varying topics
M1= adding one topic to the null model

1. ATD: anomalous topic discovery Algorithm

Input:
D: test dataset {D1, D2 …Dt}
PTM: Parsimonious topic model with M where,
M: Normal topics with j indexing= {1, 2, 3……..M}
Processing:
Step 1: identification of M0 i.e. null model
Step 2: Computation of discriminate topics from D
Step 3: Repeat (step2) and set S= ∅
Step 4: Evaluate normalized length of document d*
such as, argmind∈Dt 1/Ld l0(d)
Step 5: Search the next best document to add to the cluster,
S←S U {d*}
Step 6: Re-Optimize alternative model M1 on 'S'
Step 7: compute the relative change in log-likelihood under the null and alternative models
l1(d) ∀d ∈ Dt – S
Step 8: Choose d* i.e. candidate document & add it to the cluster
Step 9: Compute
Test significance of topic M+1in d*……..[algorithm 3]
Until topic M+1 is insignificant in d*
Step 10: if d*~∈ S then
Add d* to S
Step 11: After growing of S has terminated, then
Conduct another statistical test in algorithm 4
Step 12: If S is found significantly anomalous, then
Cluster is reported as detected & remove all documents in S from the test set Dt
Step 13: Repeat until S is insignificant
Output: Discovered Cluster 'S' with significance measure

2. Algorithm to Generate Bootstrap Document

Input:
d*: candidate document
Dv : No. of document in validation set
Processing:
Step 1: Compute similarity between document d and d* using Cosine similarity measure
Step 2: Find document sparcity d'.
d' =argmaxd ρd∗ (d) ∀d = 1,... Dv
Step 3: Randomly choose one of the documents from D', d' ∼ uniform (D').
Step 4: Then, from the Ld' words in document d',
randomly choose Ld∗ words with replacement.
Output: Document db={w1b,……wLd*b}

3. Algorithm for testing significance of topic M+1 in document d*

Input:
d* : candidate document
Dv: No. of document in validation set
M0: Null Model
M1: Alternative Model
Processing:
Step 1: Evaluate actual scope of new topic θˆ d∗
For b=1 to B1 do
Step 2: Generate Bootstrap document b from algorithm 2
Step 3: Identify the scope of new topic under M1
Step 4: Compute θˆ b
End for
Output: t(θˆ d∗) i.e. Significance of the new topic in candidate document d ∗

Algorithm 4: Testing Significance of 'S'
Input:
Candidate cluster 'S'

Score (Sb)
Processing:
Step 1: For b=1 to B2
Step 2: Set Sb= ∅
Step 3: for d=1 to |S| do
Step 4: Generate bootstrap documents for d using algorithm2
Step 5: Sb← Sb U {db}
Step 6: Compute score (Sb)
Step 7: End for
Step 8: Identify M0 & M0 on Sb
Step 9: Compute Score (Sb)
Step 10: end for
Output:
p-value to measure significance of the candidate cluster

VI. MATHEMATICAL MODEL

S' is the system of ATD such that
S = {I, F, O} I is the input to the system
F is system functions
O is Systems output

Table 1

I: {I1, I2, I3 },Set of input data	I1= Training Dataset (Tr) I2= Testing Dataset (Te) I3= Node (n)
F: {F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12, F13, F14, F15, F16, F17}	F1=Upload training dataset Tr F2=Apply preprocessing i.e. stemmer and stopword algorithm F3=Word Extraction F4= Apply PTM F5= Save PTM Parameters F6= Define null model M0 F7= Upload testing dataset Te F8= Apply preprocessing F9= Define candidates anomalous cluster F10= Compute M0 F11= Define M1 F12= Define word dictionary F13= Calculate BIC F14= Add topic in cluster F15= Calculate word probabilities F16= Calculate anomaly score F17= Display report
O :{ O1, O2, O3, O4, O5}	O1 = Preprocessed dataset O2= Preprocessed word O3= word probabilities O4 = Anomaly score O5=Cluster of anomalous topic

VII. EXPERIMENTAL SETUP

A java platform i.e. jdk1.7 is utilising to develop desktop application. Jre1.7. is configured with netbeans 8.0.2 IDE on windows as well as Ubuntu 15.4 O.S. MySQL 5.3 is used to store database, specifically, wamp is used in web server environment. Minimum 4GB RAM and min i3 processor is used for system development as well as testing.
Dataset:

1. Newsgroup dataset [18]: This dataset contains news from 20 different categories. It contains approximately 20,000 newsgroup documents.

Some categories are grouped together as contains more or less similar word set. Following are the word set categories:

Cat 1: comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x
Cat 2: rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey
Cat 3: sci.crypt, sci.electronics, sci.med, sci.space
Cat 4: talk.politics.misc, talk.politics.guns, talk.politics.mideast

Cat 5: talk.religion.misc, alt.atheism, soc.religion.christian
Cat 6 : , misc.forsale

VIII. RESULT TABLE AND DISCUSSION

Table 2. Comparative Analysis

Topics	Processing time with hadoop	Processing time without hadoop
2	7.685	43.478
4	17.348	83.734
6	41.033	130.19
8	84.173	191.74

Table 2 represents the comparative analysis for processing time between with hadoop and without hadoop for given number of topics. As per given readings time effciency is improved with hadoop platform as less processing time is required than without hadoop platform.

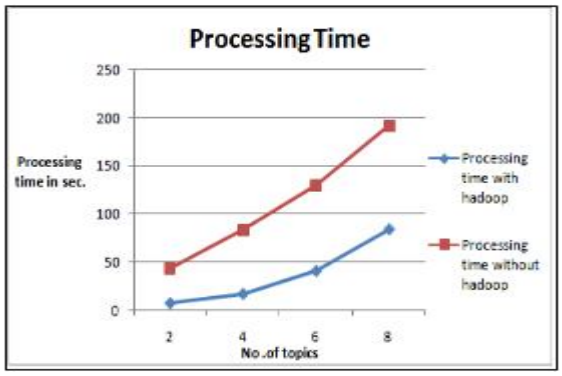

Figure 2.

Comparative analysis is depicted into figure 1 graph. Time required to process given number topics is less with with the hadoop platform than without hadoop platform.

Table 3. Precison And Recall

Total documents	Precision	Recall
174	**0.9**	**0.9**
250	**1**	**0.9**
320	**0.9**	**0.9**
450	**0.9**	**0.9**
500	**0.9**	**0.9**

In table 3, precision and recall is given for proposed system. As per observation proposed system can produced precised results with improved efficeincy.

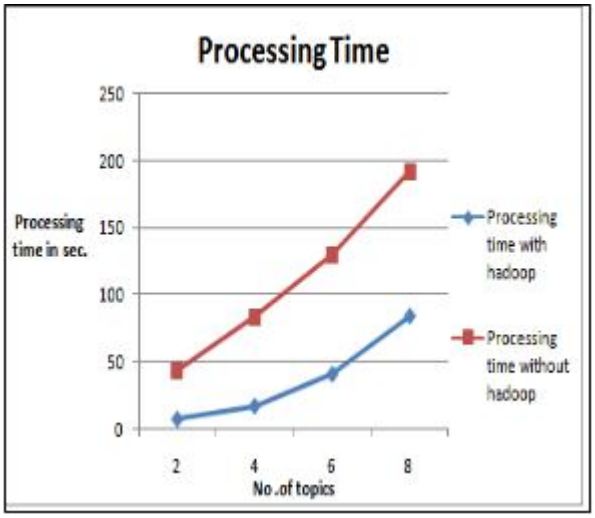

Figure 3. Graph of Precision & Recall

IX. CONCLUSION

We propose an algorithm for detecting patterns exhibited by anomalous clusters in high dimensional discrete data. 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 behavior of anomalies in group. WSARE algorithm used in rule based anomaly pattern discovery. It detects the anomaly in categorical dataset. Compared with previous system proposed system produced better results in less time and memory utilisation due to hadoop contribution. It works on synthetic as well as real datasets which can be capable of identifying group/cluster of anomalies with low density.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

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

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

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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

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

[15] 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.

[16] H. Soleimani and D. J. Miller, "Parsimonious Topic Models with Salient Word Discovery," Knowledge and Data Engineering, IEEETransaction on, vol. 27, pp. 824–837, 2015.

[17] http://qwone.com/~jason/20Newsgroups/

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