DOI 10.29042/2019-4947-4955

Helix Vol. 9 (3): 4947- 4955

Anomaly Topic and Emerging Topics Discovery Using Social Media

*[1]Yogita P. Shewale, [2]Harshal Kumar R. Khairnar

[1,] K. K. Wagh Institute of Engineering Education and Research, Nashik

[2]Mumbai Education Trust, BKC, Nashik.

Email: ypshewale@kkwagh.edu.in , harshalkhairnar27@gmail.com

Received: 13th December 2018, Accepted: 13th February 2019, Published: 30th June 2019

Abstract

Anomaly detection is an approach to detect anomalies from high dimensional discrete data. Several approaches for anomaly detection have been proposed which is only capable of detecting individual anomaly. It is very time consuming and infeasible task. With proposed ATD approach group anomalies are detected. Some techniques used all features for anomaly detection which get fail. In our system, batch of text documents are given to discover anomalies therefore, topic based algorithmic approach is utilized. With group anomalies detection, emerging topic discovery by extracting links between social users is contributed in proposed system. There is large growth in social medias detecting the latest trending topic from social medias links are receiving interest , conventional methods link text mining and text-frequency because the data is not in a social network post including images, URL'S and video so focusing on the emerging topics we required the user- links on social medias on the behavior of user which they comments on social networks basis on that be can find the anomaly that not match with the regular environment so that anomaly can comes in trend when it finds some link with recent trending environment, we calculate the anomaly score from various user which are use social medias the data set of social media may be large we need to consider social posts the datasets gathered from Facebook or twitter. The post which is consider as anomaly have time span of 30 days to be an emerging trend.

Keywords

Topic Models, Topic Discovery, Anomaly Detection, Pattern Detection

Introduction

In proposed anomaly detection system, patterns which exhibit abnormal behavior get grouped into clusters. Anomaly pattern do not tune normal behavior. AD has several applications in credit card fraud detection, insurance fraud detection, network intrusions. Traditional approaches are only capable of detecting an individual anomaly from given input. Therefore, Anomalous Topic Discovery (ATD) approach is proposed. It contains two phases such as training and testing phase. In training phase, Parsimonious Topic Model (PTM) is used. Rather than LDA model, PTM model is used to find accurate frequent words (salient words) which have accuracy & estimate the normal topics from test batch. In PTM, normal data is extracted and used to construct null model whereas, in anomaly detection phase, null model is used as, reference model to identify group or clusters of anomalies from test batch of documents. System works on training corpus to detect normal topics based on PTM model technique. Pattern matching and group anomaly in cluster is then carried out into testing phase. In testing, similar documents based on similar patterns are club into clusters hence unusual or anomalous topic remains into side. In this process, topic relevance score is calculated. In each step of proposed algorithm candidate anomalous cluster (S) is detected which exhibits maximum "deviance" from normal topic. Cluster significance is calculated to get d*. d* is candidate document belongs to S. Bootstrapping algorithm is utilized. New upcoming documents are matched with existing cluster and added to cluster having highest similarity match. If cluster size reached to limit then re-clustering is performed with specified threshold value. In experimental setup phase, performance of algorithm compared with not only semantic data set but also synthetic data set with baseline methods. To define anomalous classes a ground- truth class labels are used. In each data set, they have chosen they have chosen some classes as anomalous not match with the other classes are taken out from training and validation set. Then they normally select some documents from anomalous classes to build a test set. In this paper author do not considered any anomalous cluster actually exist in the test dataset. Document and upload time are mapped. Uploaded documents, those are anomalous but may relate to each other. System generates cluster for such topic as new trend appear with respect to time. In the propose method we consider the each post is post by users its data arrive from a social network service in proper sequential order by API every new post we consider a past interval of length T for corresponding user for training the model we propose. Step 1: we assign a score of anomaly for each post. Step 2: Aggregate that score on the basis of users by the method SDNML is a change point analysis. Step 3: Again remaining fed or provide to SDNML. Step 4 & 5: We also mention the burst-detection on kleinberg's method for change point analysis which help to find trend the post interval time.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Related Work

Application based on HMM works on credit card fraud detection as well as works on an assumption. It creates training dataset by observing the user behavior. By analyzing training dataset it defines threshold. If upcoming test record has lower value than threshold then it generates an alert as fraud detected. In this fraud detection system user spending behavior is analyzed. Based on transaction history user profile is categorized in 3 categories as: low, medium and high. According to the profile and transaction history threshold is defined. Fraud is nothing but a anomalous entry is in transaction is identified in this system [1]. PTM is parsimonious topic model used for frequent text detection in Latent Dirichlet Allocation (LDA), the words are modeled topic specific but the limitation is that if even many words occurs with similar frequencies across different topics. PTM model gives salient words which represent topics for each document to determine subset of relevant topics. BIC is finds the correctness and complexity of fit BIC is minimized to determine entire model. Results are carried out on three text corpora and an image dataset to show that proposed model can achieve higher test set. The proposed form of BIC has different parameters BIC check for different sample size to find share features increase the sample size for different parameter types in their PTM model. The use of a shared feature representation essentially increases the sample size to feature dimension ratio [2]. A rule based anomaly detection algorithm identifies anomalous topic by some predefined rules. It mainly works on emergency dataset containing emergency cases of anomalous pattern. It defines new algorithm WSARE based on the strategy: *"What's strange about recent events"*. In WSARE base method is used for computational issues the local anomaly detector detects the individual record with anomalous attribute and finds the similar pattern that have high records than the expectations it able to find the anomalous pattern in synthetic data set like hospital and shipping and network intrusion . The anomalous topic detection is at right angles statistically independent to detect local anomaly. The new event discovery and event tracing within flow of news stories publishing at subsequent stories so that it represent based on miss and false alarm rates. In this bootstrap method is used for topic detection the TDT tasks and DARPA both evaluate the tracking event a group found in task that "State of the Art" is eligible providing adequate detection and tracking events low failure rate. In TDT task represent vertical search engine in financial field result are found in group of multiple topic with stock in this method clustering is called as hierarchical galaxy. Online topic discovery and tracing the galaxy hierarchical clustering method into two steps the time and final study the effect on the similarity between two stories. The SVM is single-class classification in the circumstances of information retrieval use as single class SVM . Topic detection and topic tracking with the proposed approach splits, 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 system for adaptive anomalous discovery based on adaptive AD algorithm. It works on high dimensional dataset. Ad algorithm uses score function and generates neighbor graph for n nominal values points in dataset. It identifies the points with smaller score value with respect to m topics generated by other points. The versions of SVM approximate for single-class classification in the context of information retrieval used as single- class SVM. It is robust with respect to small categories. But SVM is very sensitive to the parameter and selection on kernel. Multiple parameters are included in this proposed method involves the data representation and the decision involved in the modification on two-class method to individual class. However, it is very specific for sensitive representation and kernel in ways those which are not very specific and transparent. To measure the performance of proposed method neural network method is required. The problem of individual anomaly detection is proposed to analyze that in most of the cases anomalies are occurred in the group form. Previous methods can able to identify group of anomaly which is already present in the dataset. In this paper author proposed a hierarchical Bayes model which is known as, GLAD i.e. "Group Latent Anomaly Detection". It can accept both pair-wise & point- wise data in the form of input. This proposed model can automatically infers the group & can simultaneously detect group anomalies. In processing step, MMSB and LDA model shared the group of membership distribution for given both input. The general approach for the iterative computations of maximum-likelihood estimates the observation views incomplete data. EM algorithm is worthy of attention because of its simplicity and the generality of associated theory. Concept similar to the EM algorithm is discussed in [11] by X. Li. Meng and D. V. Duk. Their main approach is to consider statistical construction of algorithms that are simple and fast. In this paper, they discussed about intrinsic connections between EM-type algorithm and the Gibb's sampler.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

It helps to detect individual wheat from the chaff from the thousands of incoming news stream. In this paper, DPM, Discriminative Probabilistic Model is proposed. It is simple and effective topic detection model. In this paper they focused on both online and offline topic discovery using DPM. DPM does not require any complex creative models like VMF mixture and LDA. Clustering process is represented by variation of TFIDF under condition in that only valuable words are used. A bursty phenomenon of words is utilized to discover discriminative features. Furthermore, they remark DPM soft-clustering works for offline topic detection. As extend to this work author planned to explore non- Dirichlet process mixture models from topic evolution. A system for detecting group of anomalies is proposed to identify individual objects form large dataset. In some scenario group of anomalies may appear in sequential manner. It helps to identify sources or pattern of anomaly. It takes high dimensional discrete data as an input. The scope of proposed method is applied to multidimensional dataset containing only discrete features and not applied to regular text based document anomaly detection.

M. Zhao et al [14], proposed non-parametric adaptive anomaly detection algorithm. The proposed algorithm derived from nearest neighbor graphs on nominal data point. Whenever, test samples falls down anomaly score get detected. The proposed algorithm is efficient and linear in dimension as well as quadratic into data size. K- nearest neighbor taken as an input to produced sample test score. With the computation of high dimensional quantities, it is reliably difficult in high dimensional feature spaces. Computing high dimensional quantities get avoided with the computation of score functions. Computational cost of proposed algorithm is grows linearly and quadratically in the dimension and in data size respectively.

S. Wilks, et al[15], applied the principle of maximum likelihood. A method is suggested for the functions of observation which is called as, “composite statistical hypothesis”/ “simple composite hypotheses”. For test significance a number of statics are used to make significance test which expressed in terms of ʎ.

G. Schwarz [16], concentrated towards an appropriate modification of maximum likelihood. A leading terms of Bayes estimator turns into maximum likelihood estimator. Hence they have lower probability on lower- dimensional subspaces. A statistical problem of selecting an appropriate dimensionality model which fit to the given dataset.

Problem Formulation

“To design and develop a system for group anomaly and emerging topic discovery using ATD technique.”

Computing the Link-Anomaly Score:

To compute the deviation and statistical analysis on users behavior from the normal behavior modeled. To compute the anomaly score of x =(t,u,k,v) a new comment by user (u) at time t and containing k which mentions to users To compute the probability =V , with the training set T (t)u ,which is the collection of Comment (post) by user u in the time period [t-T,u](use time interval T = 30 days).

Burst rate for the link-anomaly score is defined as follows:

$$s(\boldsymbol{x}) = -\log\left(P\left(k \mid \mathcal{T}_u^{(t)}\right)\prod_{v\in V} P\left(v \mid \mathcal{T}_u^{(t)}\right)\right)$$
$$= -\log P\left(k \mid \mathcal{T}_u^{(t)}\right) - \sum_{v\in V}\log P\left(v \mid \mathcal{T}_u^{(t)}\right).$$

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

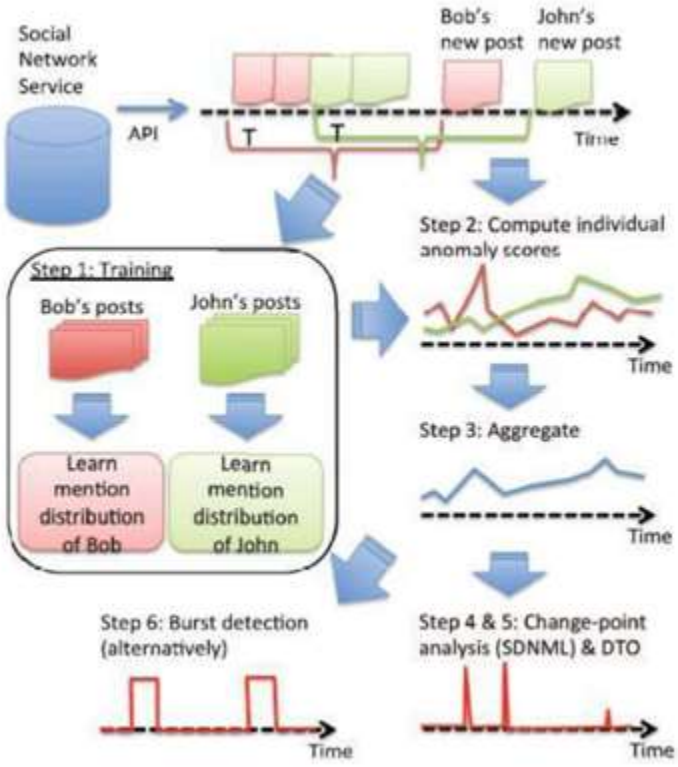


Figure 1: Overall Working Flow of Proposed Method

System Architecture

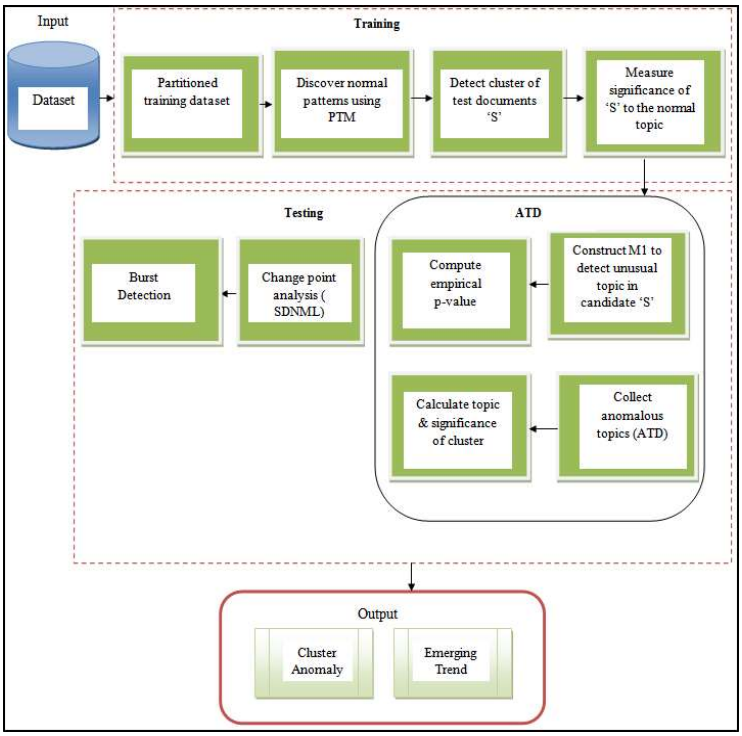


Figure 2: System Architecture

Figure 2 represents the architecture of proposed system. Our proposed ATD algorithm is to detect cluster of anomalies.

- Anomaly detection is an approach to detect anomalies from high dimensional discrete data. Several approaches for anomaly detection have been proposed which is only capable of detecting individual anomaly. It is very

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

time consuming and infeasible task. Therefore, proposed system (ATD) aims to detect group of anomalies.

- Batch of text documents are given to discover anomalies and due to topic based approach proposed algorithm can efficiently discovers topics in text documents.

A. Training Phase:

1. Upload training dataset
2. Apply stopword algorithm
3. Apply Stemming
4. Apply PTM
5. Save PTM parameters
6. Define Null model M0

B. Testing Phase:

1. Define testing dataset
2. Apply stopword algorithm
3. Apply Stemming
4. Define candidate anomalous cluster
5. Define M0
6. Define M1
7. Define word dictionary
8. Calculate BIC
9. Calculate degree of deviation
10. Calculate anomaly score
11. Add topic in cluster
12. Calculate word probabilities
13. Apply SDNML
14. Burst Detection
15. Calculate Anomaly score
16. View analysis report i.e. cluster anomaly and emerging trend

Algorithms

Input:

- Test dataset D and PTM model with 'M'

Where, D: set of documents indexed by d ε {1, 2 …D} and **'M':** Normal topics

- Link between social user's

Output:

- Detect cluster S with significant Score value measure p-value(S).
- Current trend

Processing steps:

1. READ Test dataset Dt and M0 = {θ0, H0} on Dt Where, M0 is null model,
Where,
'θ0': Topic-specific word probabilities 'H0': Topic proportions
2. COMPUTE l0 (d) ∀d ε Dt
Where, l0 (d) : length of document
3. REPEAT
SET S=Ø **Where,**
S : Set of normalized topics
4. SELECT d*= argmind∈Dt 1/Ld l0(d)
5. REPEAT
6. SET S S U {d*}
Where, d* : selected documents from 'S'
7. READ M1={θ1,H1} on S
Where, M1 : alternative model
8. COMPUTE l1(d) ∀dε Dt-S
9. SELECT
10. CALL algorithm (3), to test significance of topic M+1 in document d*
11. UNTIL if d* is insignificant topic M+1 in d*
12. EVALUATE score(S)
13. CALL algorithm (4), to test significance of 'S'

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

14. Dt Dt – S

15. TILL 'S' is significant

$$s(\boldsymbol{x}) = -\log\left(P\left(k \mid \mathcal{T}_u^{(t)}\right)\prod_{v\in V}P\left(v \mid \mathcal{T}_u^{(t)}\right)\right)$$
$$= -\log P\left(k \mid \mathcal{T}_u^{(t)}\right) - \sum_{v\in V}\log P\left(v \mid \mathcal{T}_u^{(t)}\right).$$

16. Combining Anomaly Scores from Different Users, Dynamic Threshold Optimization (DTO)

Threshold optimization: Let l be the least index such that $\sum_{h=1}^{l} q^{(j)}(h) \geq 1-\rho$. The threshold at time j is given as

$$\eta(j) = a + \frac{b-a}{N_H - 2}(l+1).$$

Alarm output: Raise an alarm if $Score_j \geq \eta(j)$.

17. Apply Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding

18. Apply the change point analysis (Kleinberg's burst detection method) as formula,

$$P^{b}sw(1\text{-}Psw)^{n-b}\,\Pi^{n}t=1\ fexp(xt\ ;\ \alpha it)$$

Where,

- *p*sw is use to find the state transition probability,
- *b* is use for number of sequential state transitions .
- *it* (*t* =1, . . . , *n*),
- fexp(x ; α) for exponential distribution is the probability density function of the rate parameter α,
- *xt* is the inter-event interval t.

19. GET emerging trend(dataset using video, image, Audio)

2. Algorithm to Generate Bootstrap Document

Input:

d*: candidate document

Dv : No. of document in validation set

Processing:

Step 1: Calculate the 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: In any sequence 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.

Where, Ld: length of document

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 ∗

Where, θˆ : Significance of new topic

Algorithm 4: Testing Significance of 'S'

Input:

Validation set (Candidate cluster 'S') Score (Sb)

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

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

Experimental Setup

We have developed a desktop application using java-Jdk1.7. Mysql 5.3 is used to store database. Core i3 machine with 4GB ram is used for development and testing. Netbeans-8.0.1 IDE is used to build and test the system using Junit.

Dataset:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Figure 3: Image of Dataset

Newsgroup dataset : It contains 20 different news topics with news article. It contains approximately 20,000 newsgroup documents

1. "Youtube" Data Set:

"Youtube " data set is some of the videos that were hack or leak by the Japan Coastal Guard officer that videos were confidential .The method of the keyword is name as "Senkaku". In that, the change detection and burst detection results in link anomaly base. There were some of the post leaked while the video is hack that interval of time.

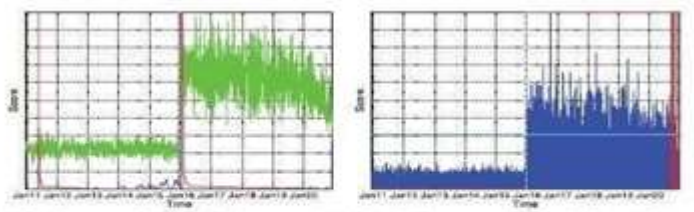
At first is 08:44 , Nov 05 , after the 9 horse the first post was leaked that is calculated and evaluated the anomaly score. Midnight, Nov 05, it found that SDNML fails. to detect this elevation as a change-point. In fact, the link anomaly-based burst detection raised an alarm at 00:07, which is earlier than the keyword-frequency-based dynamic thresholding and closer to the keyword- frequency based burst detection.

2. "NASA" Data Set

The alarm time of the text-anomaly-based burst detection was 01:24, Nov 05. 5.4 "NASA" Data Set.

This data set is related to the discussion among Twitter users interested in astronomy that preceded NASA's press conference about discovery of an arsenic-eating organism.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495



(a) Link-anomaly-based change-point analysis. Green: Aggregated anomaly score (8) at $\tau = 10$ minutes. Blue: Change-point score (9). Red: Alarm time.

(b) Link-anomaly-based burst detection. Blue: Aggregated anomaly score (8) at $\tau = 1$ second. Cyan: threshold for the filtering step in Kleinberg's burst model. Red: Burst state.

Figure 4: Result of Synthetic Data Set

Result Tables:

Total documents	Anomalous Documents (Existing System)	Anomalous Documents (Proposed System)
50	13	5
100	27	12
150	43	30
200	60	25

Table 1: Comparative Analysis of Anomaly Detection

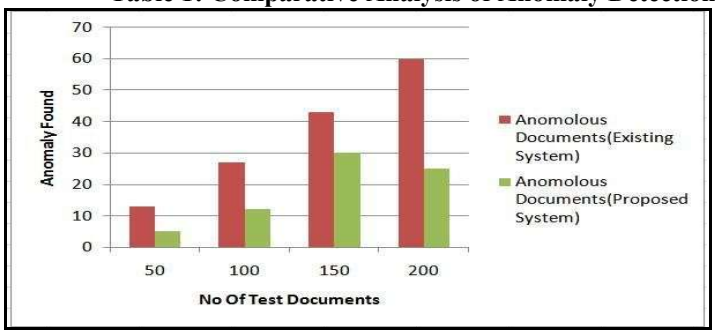


Figure 5: Graph of System Comparison

No. Of Documents	Avg. Significance	Avg. Significance
50	0.141286828	0.063787304
100	0.151248	0.08652
150	0.14982	0.09236

Table 2: Average Significance of ATD

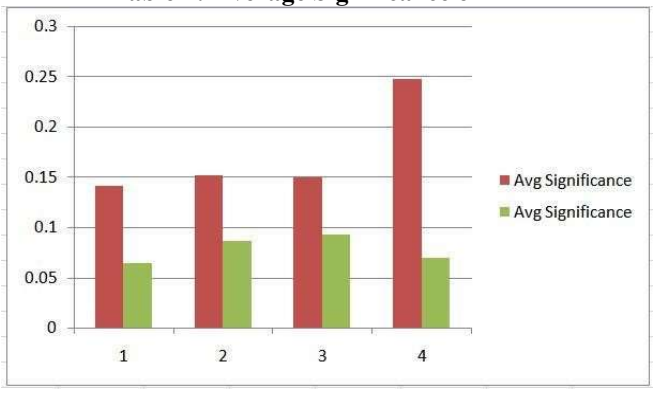


Figure 6: Graph of Average Significance

Conclusion

We proposed ATD approach to detect cluster of anomalies from input dataset. Traditional approaches of anomaly detection such as, MGMM and FGM can efficiently works on high density dataset. But it can only detect individual anomaly from huge data input which is infeasible task. Hence, proposed approach mainly aims to discover group anomaly. PTM model is utilised for normal topic discovery in training phase whereas, in testing phase, it is used to construct M1. Anomalies are nothing but abnormal patterns, in cluster formation relevance score is used for construct anomaly cluster's. With proposed work system contributes emerging trend detection. With experimental set up, proposed system proves it's efficiency in terms of accuracy.

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] M. Zhao and V. Saligrama, "Anomaly Detection with Score functions based on Nearest Neighbor Graphs," in Advances

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495