Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

Data Mining Approach for Anomaly Detection in Social Network Analysis

M.Swarna Sudha K.Arun Priya, A.Kanaka Lakshmi, A.Kruthika, D.Lakshmi Priya
Department of CSE
Ramco Institute of Technology
Rajapalayam, Tamil Nadu ,India
swarna@ritrjpm.ac.in

Dr. Valarmathi K
Department of ECE,
P.S.R.ENGINEERING COLLEGE
Sevalpatti, Sivakasi Tamil Nadu,India
krvalarmathi@yahoo.co.in

Abstract: ***Nowadays, users are more addicted to the Online Social Networks (OSN's), a network in which many users, group of people, or a particular organization are Connected via network. Use of online Social Networks has exploded and thus, causing a need of studying and understanding user behavior. Several approaches have studied the identification of anomaly detection In this paper, We propose an efficient method for anomaly detection from social networks. The present study aims at detecting the abnormal activities exhibiting different behaviors in social media application using Behavior-based anomaly detection approach. Anomalous users are detected based on their behavioral dissimilarity from others. A rich feature set is proposed for outlier detection. A method for providing visual explanation for the results is also proposed. This analysis is carried out using Facebook dataset.***

Keywords Online Social Network, Data Mining, Anomaly Detection, user's social behavior,

I. INTRODUCTION

With the advancements of technology there comes a great need of studying user behavior in social networks. With an increase in the use and benefits of OSNs comes an increase in various challenges. One of the major challenges facing such networks today is the anomaly detection. The user can be detected by analyzing the user's behavior. Various techniques are used to determine the user's behavior. The existing system includes the profile analysis method of differentiating behavior on the basis of keep tracking user's introversive behavior and extroversive behavior Online Social Network(OSN) provides sharing of one's views, ideas with others, uploading their photos, sending message in order to build connections, also can be able to see the other user's latest updates who are in connections. The existing approaches includes the method that analyzes the user online behavior by profiling user's social behavior so that the hacker can be easily differentiated from the original one. Profiling user's social behavior includes the analysis of user's updates, type of messages send by the user, click streams done by the user, types of photos uploads in the account, posts made in the account, comment made by the user on others post, etc. Our paper proposes a solution to predict the abnormal behavior of the user whose behavior diverges from the normal user. An anomaly is a set of activities that deviate from the normal behavior of the user. Anomalies are also called as outliers, abnormities. Anomaly detection is the processing technique for removing anomalous data from the dataset. Therefore, in this paper technique to analyze and detect the anomalous behavior is covered. The paper is organized into various sections. In section II data mining topic is discussed. In section III related works are discussed. In section IV data mining approaches to anomaly detection are discussed. Section V will discuss the anomaly detection techniques available. Section VI presents the risk assessment based on user behavior. Section VII risky behaviors in online social networks are discussed. Section VIII tells how the system is implemented. Section IX Clustering approaches is discussed. Section X conclusion for the paper is discussed.

II. DATA MINING

Data mining is the process of discovering the patterns, associations, changes, anomalies from large amounts of data stored in datasets. The discovery process consists of an iterative sequence of the following steps:

Data cleaning – This handles noisy, erroneous, missing or irrelevant data from the collection.

Data integration – In this stage multiple, heterogeneous are combined in a common source.

Data selection – The data relevant to the analysis task are retrieved from the dataset.

Data transformation – Where the data are transformed are consolidated into forms appropriate for mining by performing aggregations operations.

Data mining – It is a crucial step, where intelligent methods are applied to extract data patterns.

Pattern evaluation – In this step, interesting patterns representing knowledge are identified based on given measures.

Knowledge representation – It is a final phase, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

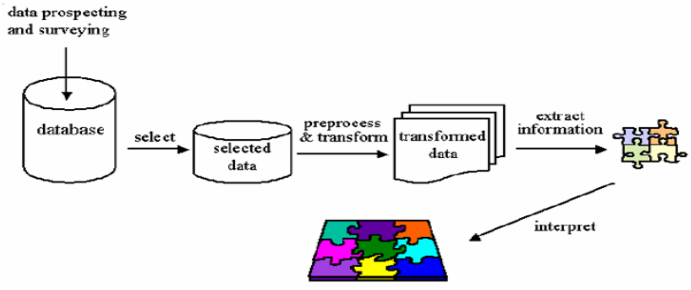


Fig II: Data mining process

III. RELATED WORK

Anomaly detection is a mature area of research. It has been successfully applied to various areas like network intrusion detection, credit-card fraud detection, email based network analysis etc. Anomaly detection algorithms work on the premise that normal behavior is more pre-dominant than abnormal behavior and will be exhibited by majority of the network entities. Chandola et al. [3] presents a comprehensive review of this. Wasserman and Faust [4] proposed theuse of centrality based measures for social network analysis. Lin and Chalupsky [5] had proposed the use of indirect connections to detect anomalous subscribers from mail data records. Chalupsky [6] used it to solve the VAST 2008 challenge [8]. While previous research work mostly views it as a classification task, we treat it as an anomaly detection problem. The false rumors are viewed as anomalies and we perform factor analysis of mixed data on our proposed features to detect these anomalies. Two strategies based on Euclidean distance and cosine similarity are proposed to describe the deviation degree. A rank based on deviation degree is computed which can facilitate further rumor detection. We show our method can achieve good performance and can shed light on automatic detection of false rumors on online social networks.

IV. DATA MINING APPROACHES TO ANOMALY DETECTION

Anomaly detection is defined as "an observation that deviates so much from the other observation". From data mining perspectives, anomaly detection is classified into the following categories:

- Supervised method
- Semi-supervised method
- Unsupervised method

A) SUPERVISED METHOD

It involves studying anomaly detection as a classification problem with the pre-labeled data, labeled either as normal or as anomalous. These methods model both the normal and abnormal behaviors. There are two applicable approaches for it.

- One, experts may pre-label the normal data and any such data which is not analogous to this model is considered anomalous.
-

The other way is to do the opposite i.e. have the predefined set of anomalous data and any objects not corresponding to the set of anomalous data are considered as normal

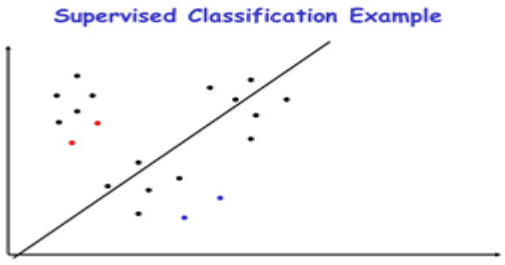


Fig IV: A(1) Supervised method

B) SEMI-SUPERVISED METHOD

Semi-supervised method, where the model is trained using the normal data, only to build the profile of normal activity. Semi-supervised methods work with two sets of data, labeled and unlabeled. So, these methods are used when out of the complete data set only few instances of data labeled as normal are available. Using the small amount of labeled data a classifier can be constructed which then tries to label the unlabeled data. Hence, a model for normal data objects is built which is used to detect the anomalies in a way that the objects not fitting the normal model are classified as anomalies. This is the simplest approach called self-training used under semi-supervised model.

Fig IV: B(1) Semi-supervised method

C) UNSUPERVISED METHOD

Unsupervised methods, where the anomaly detection model is trained using unlabeled data that consists of both normal as well as abnormal data. Unsupervised anomaly detection methods are used when labeled data objects are not available i.e. no predefined labels as "anomalies" or "normal" are attached to the data objects. Unsupervised methods are usually studied as a clustering problem. Finding the group of objects such that the objects in the group will be similar to one another and different from the objects in other group. Two major challenges faced by unsupervised methods are as follows:

- First, a data object not belonging to any cluster is considered as anomalous but many times this deliberation can be false because, such a data object can be a noise rather than an anomaly.
- Secondly, what is usually practiced is to firstly find the clusters and then the anomalies. But this methodology seems to be quite expensive as number of anomalies present in a data set is pretty less than normal data objects.

Fig IV: C(1) Unsupervised method

V. ANOMALY DETECTION TECHNIQUES

In social networks, the node disobeying these similarity measures by following behavior which is deviates from the other nodes are detect as anomalous. Anomaly detection techniques in social networks can be categorized as follows:

- Behavior based techniques
- Structural (graph) based techniques
- Spectral based techniques

A) BEHAVIOR BASED TECHNIQUES

Behavior based techniques handle the behavioral properties of the users such as number and content of message, content of the items shared, number of likes or comments on a post and duration of a conversation., shared, number of likes or comments on a post and duration of a conversation.

B) STRUCTURAL BASED TECHNIQUES

Structural based methods work on the basic principle of using structural properties to check the characteristics of normal and anomalous users. A particular graph metric is figured out for different nodes and structure and the nodes showing different values than other users are considered as anomalous. Just like supervised approach, here also a predefined normal pattern is already known and any deviation from that known pattern depicts the anomalous behavior. These techniques help to identify dynamic unlabeled anomalies by predicting future events and analyzing previous network behavior which is a precondition for dynamic anomalies.

C) SPECTRAL BASED TECHNIQUES

Spectral anomaly detection techniques help in detecting anomalies using some spectral characteristic in the spectral space of the graph.

VI. RISK ASSESSMENT BASED ON USER BEHAVIOR

Our main goal is to associate a risk score with the user based how he/she behaves in online social networks. The key idea is that the more the user behavior diverges from what is considered to be a normal behavior, more it should considered being risky. Therefore we first define a user behavioral profile.

In an online social network variety of activity are possible, such as post, comment, share, like and different types of interaction. In designing this behavioral profile, there is no need for monitoring all the users' activities. But only those that might reveal risky behavior is monitored. For example, writing more comments/posts without receiving any likes on them. Can we consider as a warning and they might be victim of an attack. On the contrary, having huge number of friends, posts, comments, likes in a short period is also consider as risky. The online social network population is heterogeneous in observed behavior, so it is not possible to define unique standard behavioral model that fits all online social network user behaviors.

However, we expect similar people tend to follow similar rules which results in similar behavioral model. For this reason, we propose risk assessment organized into two phases: The first phase aims at identifying the groups in online social network. This is achieved by clustering techniques. The second phase aims at creating behavioral model for each group identified by first phase.

VII. RISKY BEHAVIORS IN ONLINE SOCIAL NETWORKS

Due to the popularity of social network sites, cyber criminals or attackers started to exploit them in propagating malwares and carry out scams. The discrepancy is defined in terms of frequency, number, as well as type of activities.

In general, attackers use online social network infrastructure to collect and expose personal information about a user and their friends by making the user to click on some specific malicious links so as to propagate them in the network. In the following, we illustrate the most notable types of attacks:

Sybil attack: Sybil attack is one of the practical attacks in online social networks. As an example there are more than thousand social bots has been detected on Facebook. To launch this attack, a malicious user has to create multiple fake identities known as Sybil's, with the purpose to legitimate his/her identity. After that, the attackers initiate their work by sending friend requests to users in the community. Once the request has been accepted, the social bot can gather the user's private data.

Identity clone attack: In this attack, malicious user creates similar or identical profiles. The key goal is to obtain personal information about the victim's friend after successfully forging the victim and established an increased level of trust with the victim's social circle. Afterwards, he/she sends friend requests to the victim's contacts. Once the friend requests have been accepted, he builds the victim's friend network and gains access to the profiles of the victim's friends.

Socware attacks: In this attack, an adversary creates malware items, called socware, in the form of applications, pages or events containing malicious link to be propagated in the online social networks. This attack lures victims by offering false rewards to who will install/accept the socware. Once users have installed the socware, it is not only gets access to the user's personal information but also gains ability to post them on the victim's wall. As a consequence the users unknowingly end up sending socware messages or post to their friends, essentially assisting the socware's viral spread.

VIII. METHODOLOGY

Our Methodology flows through the concepts of the collection of the dataset, removing the null values, categorizing the data, clustering, risk analysis.

1. A dataset is a collection of data corresponds to the content of single database table.
2. From the dataset remove the null values.
3. Risk assessment approach is based on the idea of estimating the user's risk on the basis of how much his\her behavior deviates from the normal user. Risk assessment is composed of two phase clustering :
    a. The first phase consists of organizing the users according to group identification features.
    b. In the second phase, users are categorized using behavioral features using K-means and Expectation Maximization algorithm.

4. The risk score is estimated using Membership probability based on a value of group identification features.

   a. High membership probability-normal user.

   b. Low membership probability-abnormal user.

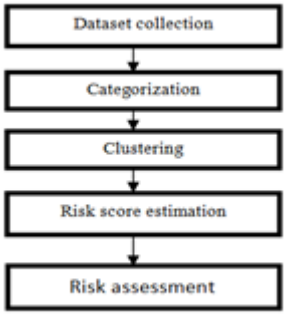
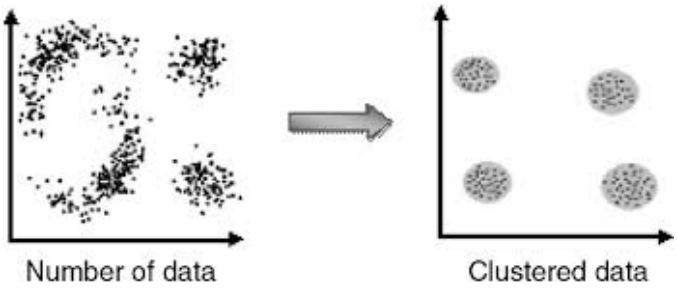


Fig VIII: Methodology

IX. CLUSTERING

Clustering is a task of partitioning a set of data based on their characteristics into clusters.It is used for pattern recognition; image processing and data analysis. It discovers structures and patterns in high dimensional data.



A) K-Means Clustering Algorithm

- K-Means is one of the simplest unsupervised learning algorithm to cluster n objects based on attributes into k partitions where k>n.
- The goal is to find groups in data, with the number of groups represented by variable K.

Step1 - **Initialization:** Randomly choose k vectors from the dataset and make them initial cluster centers.
Step2 - **Assignment:** Assign each vector to its closest center.
Step3 - **Updating:** Replace each center with a mean of its members.
Step4 - **Iteration:** Repeat step 2 and 3, until there is no more updating.

B) **Anomaly detection**
The anomaly detection algorithm proposed in this paper is based on the principle of outlier detection. An outlier in a dataset is defined as an observation which deviates so much from other observations so as to arouse suspicions that it was generated by a different mechanism.

In most real scenarios outliers are outnumbered heavily by the more commonly occurring observations. Unsupervised approaches to identify outliers are useful in detecting anomalies or nonconforming entities from large datasets. In this paper, an outlier is defined as follows [7]: Definition: Outliers of a set are the top n data elements that are farthest from their kth nearest neighbours.

IX.B (1) BOX PLOT

It is method of representing statistical data on a plot in rectangle for depicting groups of numerical data through their quartiles and vertical line inside to indicate the median value.

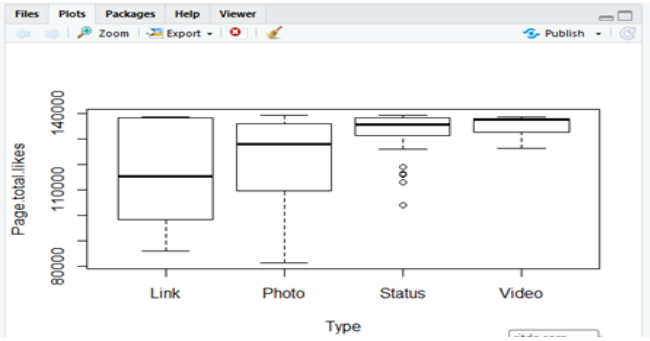


Fig IX: A(1) Box plot

IX.B (2) OUTLIERS

It is an observation point that is distant from other observations. It may indicate experimental error or variability in the measurement. It can cause problem in statistical analysis.

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

IX.B (3) HISTOGRAM

A histogram is a type of graph that is widely used in mathematics, especially in statistics. The histogram represents the frequency of occurrence of specific phenomena which lie within a specific range of values, which are arranged in consecutive and fixed intervals. The frequency of the data occurrence is represented by a bar It is an accurate representation of distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.

X EXPERIMENTAL RESULTS

The proposed approach has been tested on considering the total interaction column in the face book dataset using R programming. The Maximum and minimum deviations of Total interaction column has been identified by Histogram. Then minimally deviated values are removed as outliers using Box plot. The Box plot graph should be validated in such a way that all the outliers are removed .Now ,the Histogram has been drawn for the current data will distributed in the bars eventually.

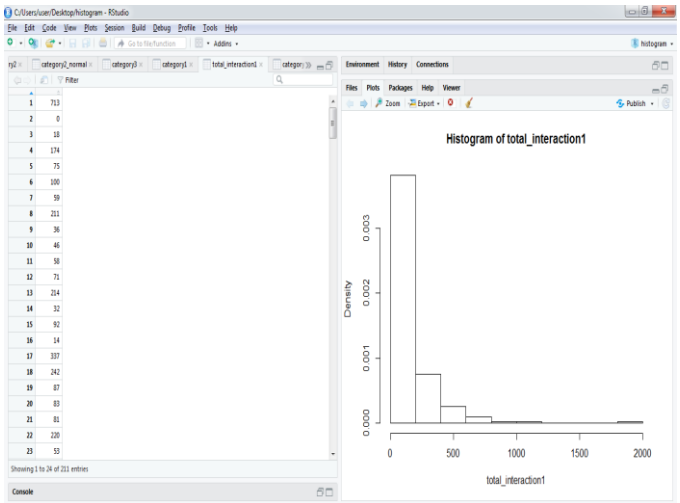


Fig : X(1) Histogram for Total interaction column

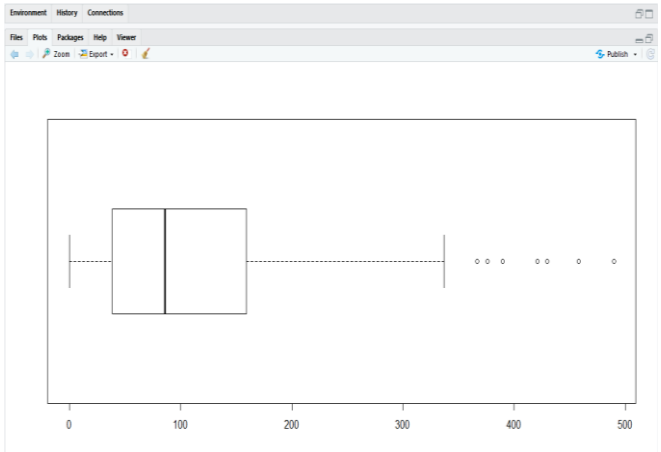


Fig : X(2) Box plot to identify the outliers

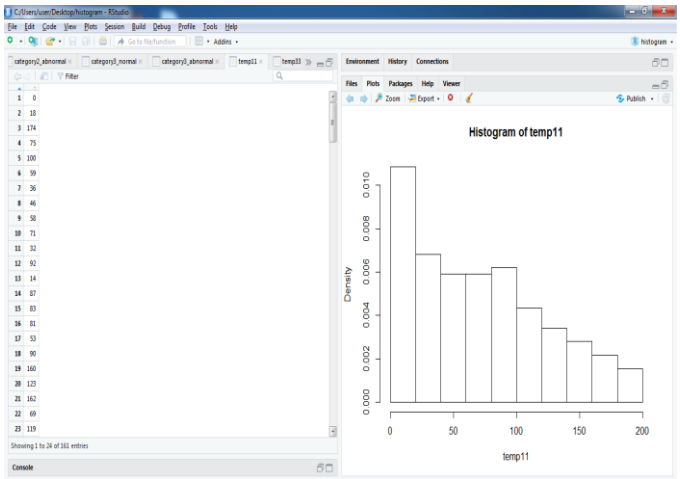


Fig : X(3) Histogram for the same Total interaction column with removed outliers

.

XI. CONCLUSION

In this paper, we proposed the method to detect the anomalous users based on their behavioral dissimilarity from others by using the outlier detection. A method for providing visual explanation for the results is also proposed. This analysis is carried out using Facebook dataset assign risk score to each OSN user. This assessment is based on user behavior under idea more user behavior diverge, it is considered to be risky. This analysis is carried out in Facebook dataset for effectiveness of our estimation

REFERENCES

[1] Naeimeh Laleh, Barbara Carminati and Elena Ferrari. Risk Assessment in Social Networks based on User Anomalous Behaviours.2014.

[2] Ravneet kaur, Sarbjeet Singh. A survey of data mining and social network analysis based on anomaly detection techniques.2015

[3]. V. Chandola, A. Banerjee, V. Kumar.: Anomaly Detection: A Survey *ACM Computing Surveys* (2009)

[4] Flora Amato, Giovanni Cozzolino, Antonino Mazzeo, and Sara Romano. Detecting anomalies in twitter stream for public security issues.2016.

[5]. S. Wasserman, K. Faust*: Social Network Analysis: Methods & Applications*. Cambridge, UK: Cambridge University Press (1994)

[6]. H. Chalupsky: Using KOJAK Link Discovery Tools to Solve the Cell Phone Calls Mini Challenge. *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, Portugal (2008)

[7]Shan-Hung Wu, Man-Ju Chou, Chun-Hsiung Tseng, Yuh- Jye Lee, and Kuan-Ta Chen, Senior Member, IEEE

8]. IEEE VAST Challenge 2008, http://www.cs.umd.edu/hcil/VASTchallenge08

[9]. S. Lin and H. Chalupsky: Discovering and explaining abnormal nodes in semantic graphs. *IEEE Transactions on Knowledge and Data Engineering*, Volume 20 (2008)

[10] Naeimeh Laleh, Barbara Carminati and Elena Ferrari. Anomalous change detection in time-evolving OSNs.2016.

[11] Renjun Hut, Charu C.Aggarwal, Shuai Mat and Jinpeng Huait.An embedding

.