Mini Project Report

On

Anomalies Detection in Social Network

Submitted by

182286
Nishant Raj
182IS012
M-Tech CS-IS (1st Semester)

Guided by

Dr. M. Venkatesan

Department of Computer Science and Engineering,
NITK, Surathkal

Date of Submission: 20-11-2018

Department of Computer Science and Engineering
National Institute of Technology Karnataka, Surathkal.
2018-2019

DECLARATIONS

I hereby declared that the work, which is being presented in the report entitled "**ANOMALIES DETECTION IN SOCIAL NETWORK**" submitted in the requirement for the degree of Master of Technology in **Computer Science & Engineering – Information Security,** submitted in the department of **Computer Science & Engineering** at National Institute of Technology Karnataka, Surathkal is an authentic record of my own work carried under the supervision of **Dr. M. Venkatesan**. I have not submitted the matter embodied in this report for any other degree.

Nishant Raj
Department of Computer Science and Engineering
(182IS012)

Dr. M. Venkatesan
Guide

CERTIFICATE

This is to certify that the Mini Project-I Semester Report entitled "**ANOMALIES DETECTION IN SOCIAL NETWORK**" submitted by Nishant Raj (182IS012), as the record of the work carried out by him, is accepted as the Mini Project-I Report submission in partial fulfilment of the requirements for the award of degree of Master of Technology in Computer Science & Engineering by National Institute of Technology Karnataka, Surathkal for session 2018-2019.

Dr. M. Venkatesan

Guide

ACKNOWLEDGEMENT

I have taken efforts in this report work. However, it would not have been possible without the kind support and help of many individuals. We would like to extend sincere thanks to all of them.

I would like to acknowledge the kind support rendered by my supervisor Dr. M. Venkatesan for conducting the Mini Project.

I am highly indebted to Department of Computer Science & Engineering and my institute National Institute of Technology Karnataka, Surathkal for providing me this opportunity of gaining in depth knowledge of such a trending topic of research.

Nishant Raj

(182IS012)

Abstract

Author:	Nishant Raj	
Name of Study:	Anomalies detection using Random Forest Algorithm	
Date:	19.11.2018	**Pages:** 13

Data mining is the way of extracting the useful information and/or patterns from large volume of information by using various techniques. It is an important area of research and is pragmatically used in different domains like finance, analysis on social network data, etc.

In the past decade, network structures have penetrated nearly every aspect of our lives. The detection of anomalous vertices in these networks has become increasingly important, such as in exposing computer network intruders or identifying fake online reviews.

This project is about anomaly detection in social network, completely on network structure. We use random forest algorithm to train and test our classifier. Also, we are going to see how the effect of increase of trees in forest to the accuracy of prediction.

Keywords: Data Mining, network structure, fake-real users, Random Forest Algorithm, Trees

Table of contents

1 Introduction

Social media systems provide convenient platforms for people to share, communicate, and collaborate. While people enjoy the openness and convenience of social media, many malicious behaviours, such as bullying, terrorist attack planning, and fraud information dissemination, can happen.

Therefore, it is extremely important that we can detect these abnormal activities as accurately and early as possible to prevent disasters and attacks. Needless to say, as more social information becomes available, the most challenging question is what useful patterns could be extracted from this influx of social media data to help with the detection task. [2]

By definition, anomaly detection aims to find "*an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*"

Mapping to social media analysis, we can recognize two major types of anomalies: [2]

- Point Anomaly: the abnormal behaviours of individual users
- Group Anomaly: the unusual patterns of groups of people

One of the most important tools used in this is Data Mining Techniques. There are many techniques the most important of all is Random Forest algorithm. Why so will be discussed in next section.

Random Forest algorithm: Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity.

If we are considering about the point anomaly it is easy to define as group in larger scale as anomaly.

2 Methodology

2.1 Literature survey

Complex networks are deined as systems in nature and society whose structure is irregular, complex, and dynamically evolving in time. The can consist of thousands, millions, or even billions of vertices and edges [6]. These systems occur in every part of our daily life, including, food webs, the Internet, and online social networks [1]. Analyzing the unique structures of such networks can be revealing; for example, an analysis of network structures can indicate how a virus will propagate most quickly in a computer network, or which vertex malfunction in a power grid will affect more houses.

Many studies have shown that vertices which deviate from normal behavior can offer important insights into a network [3]. For instance, Fire et al. [7] observed that fake profiles and bots in social networks have a higher probability than benign users of being connected to a greater number of communities.

2.1.2 Basic Methodology of anomaly detection technique[3]

Although different anomaly approaches exists, as shown in figure 1 parameter wise train a model prior to detection.

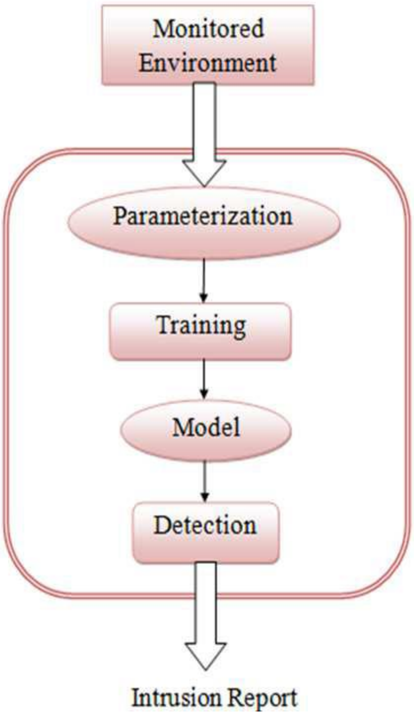


Figure 1: Methodology of Anomaly Detection

Parameterization: Pre processing data into a pre-established formats such that it is acceptable or in proper format with the targeted systems behavior.

Training stage: A model is built on the basis of normal (or abnormal) behavior of the system. There are different ways that can be opted depending on the type of anomaly detection considered. It can be both manual and automatic.

Detection stage: When the model for the system is available, it is compared with the (parameterized or the pre defined) observed traffic. If the deviation found exceeds (or is less than when in the case of abnormality models) from a pre defined threshold then an alarm will be triggered.

Anomaly Detection Using Data Mining Technique

Anomalies are pattern in the data that do not conform to a well defined normal behavior. The cause of anomaly may be a malicious activity or some kind of intrusion. This abnormal behavior found in the dataset is interesting to the analyst and this is the most important feature for anomaly detection [8].

Random Forest

A Random Forest is a classifier consisting of collection of tree-structured classifiers where independent random vectors are distributed identically and each tree cast a unit vote for the most popular class at input x. A random vector is generated which is independent of the past random vectors with same distribution and a tree is generated by using the training test [3].

Brieman [4] has proposed a randomization approach that works better with bagging or random space method. To generate each tree of random forest, following steps are followed that are described below:

- Training dataset consist of N number of records.

- Sampling of N number of records are performed randomly but with replacement.

- This sample of dataset is named as bootstrap sample.

- If this training set would consist of M number of input variables, m<<M number of inputs are selected randomly out of M and the best split on these m attributes is used to split the node.

- The value of m will remain constant during forest growing.

- The tree will be grown to the largest possible level.

There are two reasons for using Bagging approach. They are given below: [4]

- It seems that the use of bagging along with the random features generates more accurate results.

• Bagging can be used to provide ongoing estimation of generalization error as well as the estimation of strength and correlation.

As we have discussed about the reasons of using Random Forest algorithm, we will move forward to the implementation of the same for anomaly detection data mining.

Studies conducted in the past several years indicate that malicious users typically present different behavioral patterns than real users. Motivated by the difference between the observed behavioral patterns of fake profiles and real users, we developed a method to generate examples for our classifier.

The building of classifier purly based of the topology of the graph of node and edges is done.

Node: Users on social network

Edges: Link between the users on social network

Label 1: Real user

Label 0: Fake user

2.2 Process Followed

As discussed earlier here also the same procedure is used, as common process have been followed as in case of any other classifier building.

1. Feature Extraction
2. Training classifier
3. Testing classifier

In detail we are going to discuss each.

1. *Feature Extraction*

We are randomly considering 15000 vertices out of 50000 vertices for the Features extraction.

For each node i.e., user in social network we are finding the following topological data.

- average_scc: Average number of successor node a node has.
- out_degree: Total number of links out of node

- in_degree: Total number of links in to node
- bi_degree: Total number of links in to node and out also from the node
- in_degree_density: Ratio of in_degree to total degree of the graph
- out_degree_density: Ratio of out_degree to total degree of the graph

2. *Training classifier*

Using the Random Forest algorithm, we are training our classifier on features extracted from data set. The classifier can predicate between fake and real user. Random Forest have more accuracy than other.

We are using 3000 featured nodes for training of the classifier.

We are also plotting graph of accuracy with different no of estimators to apply on the Random Forest algorithm.

3. *Testing classifier*

Labelled data (Real and Fake Users) is used for testing of the classifier.

We are taking around 200 random featured labelled data for testing of our classifier. We also tried using different set of data other than the given dataset to test classifier.

We have used two different logics for the selection of training dataset:

a. A part of the training dataset.

b. Different random selected testing dataset.

This is the process of followed to built and test classifier. Now we will we discuss about *source of social network dataset*.

3 Source of Social Network Datasets

Twitter dataset is easily and vastly available, but for this project comparatively a smaller number of datasets is required, due to computation infeasibility.

Network	*Directed*	*Vertices (users)*	*Links*	*Date*	*Labels*
Twitter	Yes	5,384,160	16,011,443	2012	Yes

We have taken the dataset from online data source namely http://proj.ise.bgu.ac.il/sns/datasets/twitter.csv.gz and label of the data from http://proj.ise.bgu.ac.il/sns/datasets/twitter_fake_ids.csv

Data set is of the twitter a well-known social networking website with directed links. The data set in *twitter.csv.gz* is as *src* (Source) and *des* (Destination), means user *src* is connected to user *des*.

We also have dataset from *twitter_fake_ids.csv* which is in the format that it is src(source) and label. It means src is labelled to 0, where 0 means Fake users. So, it is a dataset of all known anomalous users.

Here in *twitter.csv.gz* dataset there are around 1,40,00,000 vertices. But due to computation infeasibility on such huge dataset, we are taking only 50000 vertices.

Similarly, the same for *twitter_fake_ids.csv* dataset, we are reducing size.

Dataset explanation *twitter.csv.gz*:

Suppose A user is connected to B user.

Attributes- src: user A is source of link.

des: user B to A is connected.

Dataset explanation *twitter_fake_ids.csv*:

Suppose user A is anomaly.

Attributes- src: user A in the network.

label: 0 -> Fake user (marked).

Now, we will show and explain results found.

4. Experiment and Result analysis

We have built a classifier that can detect anomalous users from the graph based on topology of the graph only.

As the size of the dataset is increased, we can increase the accuracy of the classifier. The dataset we used is labelled, hence the accuracy of the labelling is also one of the parameters to taken in consideration to evaluate classifier.

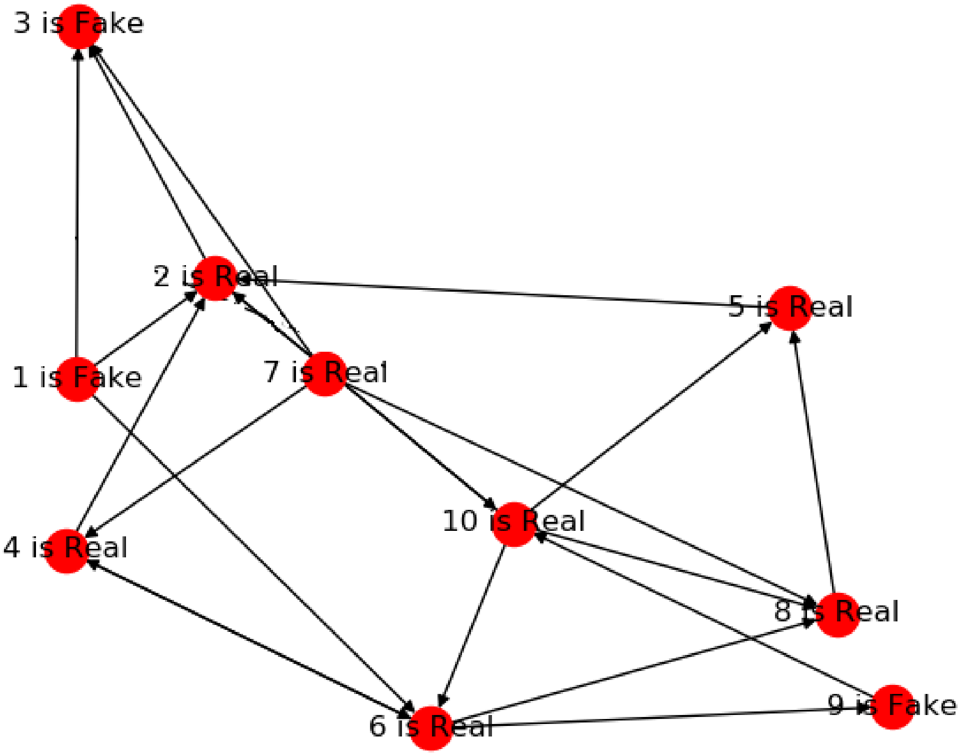


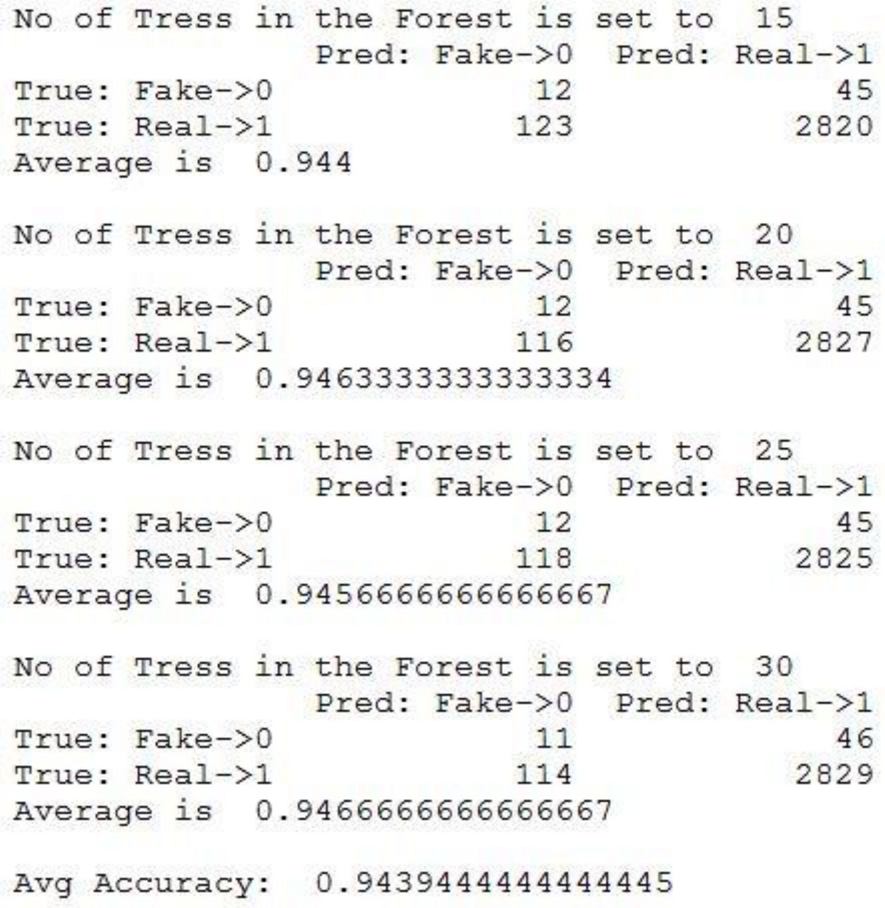
Figure 1- Simple graph showing Real and fake nodes

In figure 1. we have shown fake and real users with the link. We have just taken 10 users and 19 links into consideration.

Here, user 1,3 and 9 are Fake labelled users in network, rest are real users. We can see these users show some unusual behaviour i.e. links. User 3 doesn't have any link outward, User 1 doesn't have any inward links and user 9 has only one link in and one link out. Hence, based on these features these are marked as fake users.

Using similar type of dataset with much large size will form some criterion to classify any vertex as fake/anomalous. This is called training of classifier.

We are using same random set of dataset to verify the label given and produced by the classifier.



```
No of Tress in the Forest is set to  15
                  Pred: Fake->0  Pred: Real->1
True: Fake->0               12             45
True: Real->1              123           2820
Average is  0.944

No of Tress in the Forest is set to  20
                  Pred: Fake->0  Pred: Real->1
True: Fake->0               12             45
True: Real->1              116           2827
Average is  0.9463333333333334

No of Tress in the Forest is set to  25
                  Pred: Fake->0  Pred: Real->1
True: Fake->0               12             45
True: Real->1              118           2825
Average is  0.9456666666666667

No of Tress in the Forest is set to  30
                  Pred: Fake->0  Pred: Real->1
True: Fake->0               11             46
True: Real->1              114           2829
Average is  0.9466666666666667

Avg Accuracy:  0.9439444444444445
```

Figure 2-*Output on different forest size.*

As per the figure 2 we can see the working of the classifier.

No of trees in the Forest -> it's the number of trees used in building of Forest for the classifier training.

Prediction vs True is shown in figure. Average accuracy for each no_estimators is also shown at last of each computation. We can see the increase in number of estimators we can actually increase the accuracy.

The Total accuracy is quite high as shown in last part of the figure 2.

We also have plotted the graph of *accuracy of predication* vs *number of estimators*, and can be seen in figure 3.

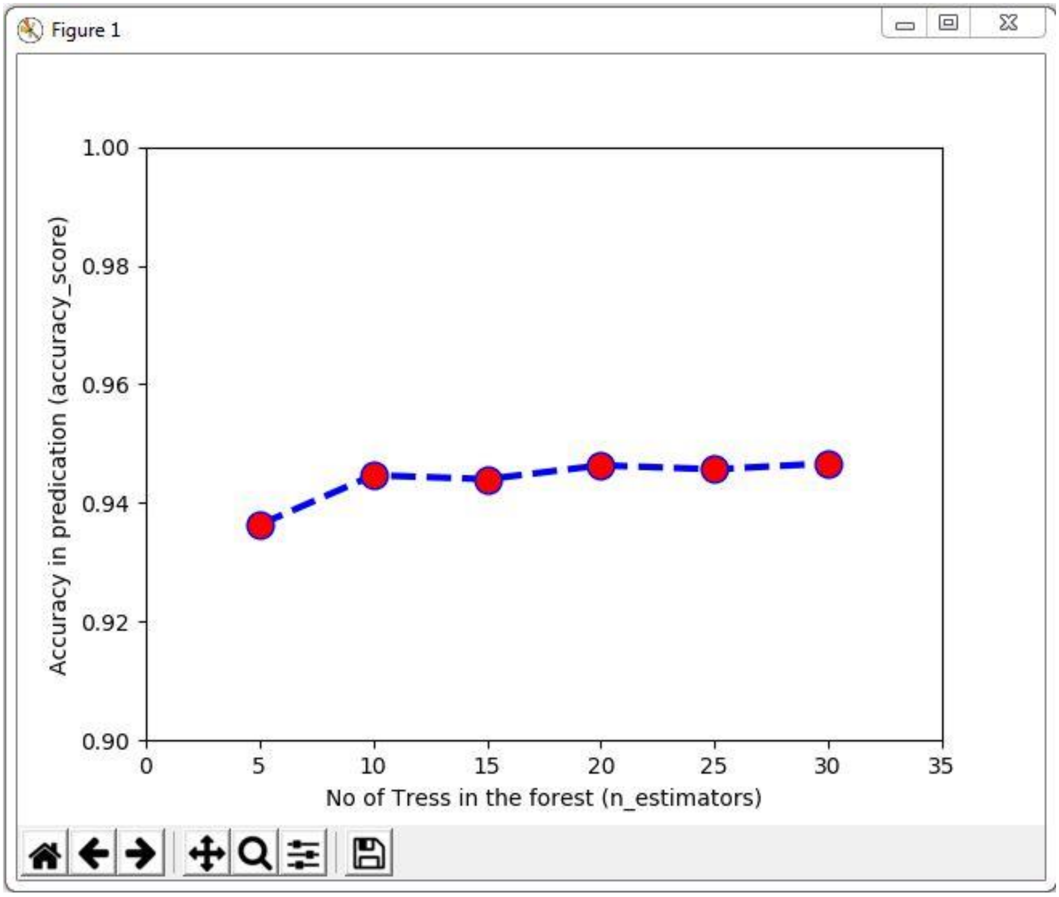


Figure 3- *Graph Accuracy in Predication vs No of Trees in the Forest*

Figure 3 is showing the graph Accuracy of prediction vs no of trees in the forest training. We can see with increase in the no of trees used in training classifier the accuracy is increased, and after some point it saturates.

This shows the capacity of our classifier if trained on Random Forest Algorithm.

5. Future Work and Conclusion

In this project we have built a classifier based on Random forest algorithm. It can classify a user to either Real or Fake. We have also shown the Accuracy of the classifier and the increase in accuracy with increase in accuracy of classifier. Random forests are an effective tool in prediction.

We can in future use different other techniques to train the classifier and have comparative study of it. We can also use more features to get better and more accurate result. Moreover, we can get more data set of different social networking website to make a universal anomaly detection classifier.

References

[1] Generic Anomalous Vertices Detection Utilizing a Link Prediction Algorithm by Dima Kagan, Yuval Eloviciy and Michael Fire, *Department of Computer Science & Engineering, University of Washington, The eScience Institute, University of Washington.*

[2] A Survey on Social Media Anomaly Detection by *Rose Yu, Huida Qiu, Department of Computer Science, University of Southern California*

[3] Information and Engineering Systems Survey on Anomaly Detection using Data Mining Techniques, *Shikha Agrawal, Jitendra Agrawal, at 19th International Conference on Knowledge Based and Intelligent.*

[4] RANDOM FORESTS, *Leo Breiman Statistics Department University of California Berkeley, CA 94720 January 2001*

[5] M. Fire, G. Katz, and Y. Elovici. Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies. *Human Journal, 1(1):26-39, 2012.*

[6] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In Proceedings of the ninth ACM SIGKDD *international conference on Knowledge discovery and data mining, pages 631-636. ACM, 2003.*

[7] M. Fire, G. Katz, and Y. Elovici. *Detecting spammers and fake profiles in social networks based on topology anomalies. Human Journal, 1(1):26-39, 2012.*

[8] Survey on Anomaly Detection using Data Mining Techniques, Shikha Agrawal, Jitendra Agrawal, Department of Computer Science and Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India. *19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems.*

[9] Lee W., Stolfo J. Salvatore, Data mining approaches for intrusion detection; Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas; 1998;p. 79-94.

[10] Random Forest: A Review Eesha Goel* , Er. Abhilasha Computer Science & Engineering &GZSCCET Bhatinda, Punjab, India. *International Journal of Advanced Research in Computer Science and Software Engineering*