THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Cluster Ensemble Methods for Detection and Classification of Malwares and Phishing Websites

Trupti Mane
Department of Computer Engineering, Mes College of Engineering, Pune, India
Sandesh Ilhe
Department of Computer Engineering, Mes College of Engineering, Pune, India
Hemant Bhaskar
Department of Computer Engineering, Mes College of Engineering, Pune, India
Akash Kamble
Department of Computer Engineering, Mes College of Engineering, Pune, India
Suraj Khade
Department of Computer Engineering, Mes College of Engineering, Pune, India

Abstract:
We are designing an automatic categorization system to automatically group phishing websites or malware samples by using a cluster ensemble by aggregating the clustering solutions generated by different base clustering algorithms. Where cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

This application will be useful to identify various internet related issues. Malwares and Phishing Web sites are one of those threats which cause harmful damages. Though Phishing Web sites are one of the new threats and look different from malwares, they possess similar characteristics. They have similarities like Driven by economic benefits, both malware and phishing websites are increasing rapidly, most of their essence is stable and etc.

Keywords: *Cluster ensemble, Malware categorization, Phishing website detection.*

1. Introduction

In today's world, Internet has become an inseparable part of our lives. Despite all the advantages, it does have a problem of Malware and Phishing websites. To overcome this and many other minor problems we present our project, which is a Cluster ensemble method for detection and classification of malwares and phishing websites. In our system we develop an automatic categorization system to automatically group phishing websites or malware samples using a cluster ensemble by aggregating the clustering solutions generated by different base clustering algorithms.

Clustering ensemble refers to the process to obtain the single and better performing clustering solution from a number of different inputs clustering for a particular dataset. There are many clustering methods available such as:-

- Hierarchical
- K mean
- K-Medoid

This system will be web based application, which ask client to choose type of testing either malware or phishing web pages. User of the system will be then allowed to upload his file to detect and classify file as malware or not. In case of phishing web page testing, user has to provide URL of website, for confirming whether web site is fake or genuine.

2. System Design

2.1. System Architecture

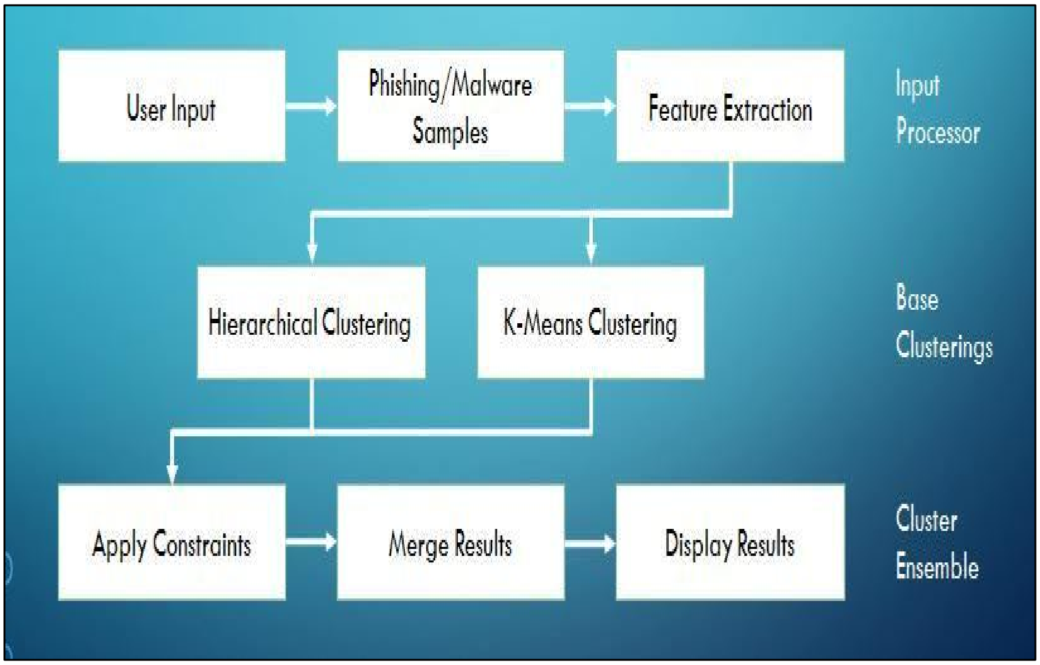

Figure 1: System Architecture

2.1.1. Input Processor

- User Input

User either tries to access a webpage or execute a program.

- Phishing/Malware Samples

If user accesses a webpage, then phishing samples are obtained from the database.
If user executes a program, then malware samples are obtained from the database.

- Feature Extraction

If user accesses a webpage, then the frequency of terms in the META, TITLE tags is found. If user accesses a webpage, then the instructions in the programs are encoded and frequency of each is found.

2.1.2. Base Clustering

- Hierarchical Clustering

Because of lower computation cost, we use the hierarchical clustering algorithm as the frame, starting with N singleton clusters, and successively merge the two nearest clusters until only one cluster remains.

- K-Means Clustering

K-Means is squared error-based partitioning clustering, which assigns a set of data points into clusters using an iterative relocation technique. A cluster is represented by the mean of closely located data points called centroid.

2.1.3. Cluster Ensemble

- Apply Constraints

The domain knowledge in the form of website-level/sample-level constraints can be naturally incorporated.

- Merge Results

Results of both base clustering are merged together to increase the accuracy of the classification of phishing website/malware.

- Display Results

Show results stating whether a website is phishing or genuine, and also shows the category of phishing website. Show results stating whether a program is malware as also the category of malware.

3. Technical Specification

3.1. Advantages

Cluster ensemble methods for automated malware and phishing web site categorization have following advantages:

- Cluster ensemble framework is applied for both, malware categorization and phishing website testing.
- Intelligent system automatically categorizes large amount of data effectively.
- Our system uses constraints provided by human analyst and merging techniques which provide more accuracy.

3.2. Disadvantages

Cluster ensemble methods for automated malware and phishing web site categorization have following disadvantages:

- System is online so user machine can get affected by malwares while he is offline.
- System should update database as quick as, instance after new threats have founded.
- Database must have large enough capacity to store new results.
- Sufficient bandwidth must be provided to serve many clients at a time.

3.3. Applications

Automated Categorization System (ACS) performs well for real phishing website categorization as well as malware categorization application. It has following applications:

- User machine stay safe from malwares.
- Users credential remains safe like bank details, emails, etc.
- Instead of checking detecting malicious programs on own machine user can check it by executing file on web based application so there are very less chance of threats for user.
- Reduced processing load for user’s machine.

4. Conclusion

We have developed an ACS (Automatic Clustering System) which can be applied for phishing website categorization as also for malware samples. We have also categorized phishing websites and malwares into families that share some common traits using an ensemble of different clustering solutions that are generated by different clustering methods. Experimental studies illustrate that our ACS system performs well for real phishing website categorization as well as malware categorization applications.

5. References

1. M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, “Synthesizing near-optimal malware specifications from suspicious behaviors,” in Proc. IEEE Symp. Secur. Priv., Washington, DC IEEE Computer Society, May 2010, pp. 45–60.
2. “Exploiting diverse observation perspectives to get insights on the malware landscape”. Corrado Leita Symantec Research Labs Sophia Antipolis, France corrado leita@symantec.com
3. M. Gheorghescu, “An automated virus classification system,” in Proc. VIRUS BULLETIN CON., Oct. 2005.
4. I. Gurrutxaga, O. Arbelaitz, J. M. Perez, J. Muguerza, J. I. Martin, and I. Perona, “Evaluation of Malware clustering based on its dynamic behaviour.,” in Proc. 7th Australas. Data Mining Conf., 2008, pp. 163–170.
5. Weiwei Zhuang, Yanfang Ye, Yong Chen, and Tao Li “Ensemble Clustering for Internet Security Application” IEEE transaction on system, man, and cybernetics- part C :Application and reviews, vol. 42, No. 6, November 2012