Pune Institute of Computer Technology
Dhankawadi, Pune

A SEMINAR REPORT
ON

SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

SUBMITTED BY

ABDUL DALVI
Roll No. 31201
Class TE 2

Under the guidance of
Prof. S.S.SHEVTEKAR

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2019-20

DEPARTMENT OF COMPUTER ENGINEERING

Pune Institute of Computer Technology

Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled

"Sentiment Analysis Using Machine Learning"

Submitted by

Abdul Dalvi Roll No. 31201

has satisfactorily completed a seminar report under the guidance of Prof. S.S.Shevtekar towards the partial fulfillment of third year Computer Engineering Semester II, Academic Year 2019-20 of Savitribai Phule Pune University.

Prof. S.S.SHEVTEKAR
Internal Guide

Prof. M.S.Takalikar
Head
Department of Computer Engineering

Place:
Date:

ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Prof. B.D.Zope and Head of Department Prof. M.S.Takalikar for their support.

I also sincerely convey my gratitude to my guide Prof. S.S.SHEVTEKAR, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement from beginning till end to make this seminar a grand success.

SAMPLE TEXT. In case you want to add something.

Contents

List of Tables

List of Figures

Abstract

Sentiment analysis deals with identifying and classifying the input source text based on sentiments and opinions. In the business field the opinions of the customers are very valuable to companies. Customers use social media platforms like twitter, or blogging web sites to express their views and opinions. Through such platforms a large amount of data is generated. This data is used as input and analysis is performed on it.

Twitter is one of the most used social networking platform, where it allows only 140 characters at a time which has a great amount of slang words, emoticons and misspellings occurrences, hence analysis becomes a difficult task. Knowledge base approach and Machine Learning approach are the two strategies used for analyzing the sentiments from texts.

We will try to emphasize on the Machine learning approach for sentiment analysis.

Keywords

Twitter, Sentiment Analysis, Machine learning.

1 INTRODUCTION

Cloud environments have nowadays evolved as the critical backbone for a number of socio-economical ICT infrastructures, due to their intrinsic capabilities such as elasticity and resource transparency. Consequently, they are becoming increasingly mission-critical since they provide always-on services for many every-day applications (e.g. IPTV), safety-critical operations, critical manufacturing services, and critical real-time services.

Cloud computing has become increasingly popular by obviating the need for users to own and maintain complex computing infrastructure. However, due to their inherent complexity and large scale, production cloud computing systems are prone to various runtime problems caused by hardware and software failures. Cloud anomaly detection is a technique of detecting anomalous behavior of network date being collected . To detect anomalies, we need to monitor the cloud execution and collect runtime performance data and network flow over cloud. These data are usually partially labeled, and thus a prior failure history is not always available in production clouds, especially for newly managed or deployed systems.

Infrastructure items, such as hosts, can be broken into by a competing company to attain confidential information about its users and other data that is stored on the machine. This in turn allows workflows to be changed, i.e. by breaking in a system and patching the code-base or the platform itself, or simply by reverse engineering workflows and creating rogue clients.

Another problem is that attacks themselves have become sneakier. Attackers tend to use more advanced techniques, and more persistence to eventually mask an attack. For example, if credentials of legitimate service users are stolen and information is leaked gradually and persistently over a longer time period. Such attacks usually manifest in a change of behavior of entities involved in any given activity (e.g. behavioural changes observed in off-key working hours, spiking access over document data etc.). To decrease the chance of successful attacks, security monitoring was introduced to analyse events committed by sensors in the corporate network. The analysis of events usually involves signature-based methods. Features, extracted from logged event data, are compared to features in attack signatures which in turn are provided by experts Other approaches, e.g. anomaly detection, often make use of machine learning-based algorithms. Anomalies are an unexpected event (or a series of unexpected events) that exhibit a significant change in behaviour of an entity, for example, a user. If anomalous behavior can be distinguished from normal behavior by hard bounds that are known beforehand, then signature-based approaches can be used to classify attacks immediately. However, when it is hard to specify all entities and their normal behaviour completely beforehand, then statistical measures have to be used to classify deviations in order to detect possible attacks.

Unfortunately, probabilities and patterns of unwanted behaviour are very hard to procure. But it is reasonable to assume that most activity in a network is not triggered by compromised machines and attacks are represented by only a tiny fraction of the overall behaviour.

Many machine learning algorithms proposes techniques to classify malwares but the true challenge lies with the fact that classification model must be dynamic in nature as the malwares are generated within seconds and its nature could be different from the generic ones which is impossible for a static analysis system to identify followed by classification during that moment.

2 MOTIVATION

With the time evolvement of time IT services and all area infrastructures are shifting to cloud services, as cloud provides such facilities of availability, storage as well computing environment. Cloud has got immense increased amount of data over it so managing that data is being a major concern over these days. Since hackers have been trying new phenomenon or procedures to get some data and malicious adding to those dataset. So machine learning is one of the solution of such a problem where we are unaware of which type of attacks exists in this field. To analyse the pattern in the dataset being used and find anomalous happenings.

In recent times successful attacks on machine learning has happened . These attacks compromise machine learning algorithms. Such compromising attacks are very sensitive whic can lead to big disasterous result

Anomaly in data is getting common so detecting anomalies is the major task. Apart from normal available patterns in data new patterns are to be detected which brings a challenge for machine learning. Efficiency of detection and dynamic detection over data is also a challenge

Thus, for effective counter measure for these poisoning attacks , to find easiest algorithm to defend through the counter attack and to make sure the effectiveness of machine learning algorithms are maintained this is needed

3 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques propose in various references:

Table 1: Literature survey

No.	Title	Description	Findings	Dataset	Limitat
1	Sentiment analyis in Twitter using Machine learning techniques	Breif study of various methods used for sentiment analysis	Machine learning techniques prove to be much efficient than symbolic techniques	Machine learning techniques require a well defined database to train on	Own dataset created
2	Comparative Analysis of sentiment orientation using SVM and Naive Bayes techniques	Sentiment analysis was performed on a movie dataset using both techniques	It was found, that drama genre had the highest accuracy rate	Opitimization and refinement required	IMDb d
3	Sentiment analysis on Movie reviews using recurrent neural network	Breif study of neural network	RNN has better accuracy than SVM and Naive bayes classifiers	Depicts result into two classes only	IMDb d
4	Deep learning for Sentiment analysis	Survey of techniques	Overview of deep learning, applications of sentiment analysis	Lacks design of models	-

4 A SURVEY ON PAPERS

4.1 A hybrid machine learning approach to network anomaly detection

The field selection from the dataset is being done by genetic algorithm in this paper. For the first operation, the method transform TCP/IP packets into binary gene strings. We convert each TCP and IP header field into a bit binary gene value, '0' or '1'. '1' means that the corresponding field exists and '0' means it does not. The initial population consists of a set of randomly generated 24-bit strings, including 13 bits for IP fields and 11 bits for TCP fields. The total number of individuals in the population should be carefully considered. If the population size is too small, then all gene chromosomes soon converge into the same gene string, making it impossible for the genetic model to generate new individuals. In contrast, if the population size is too large, then the model spends too much time calculating gene strings, negatively affecting the overall effectiveness of the method.

Two existing SVM methods: soft margin SVM and one-class SVM are introduced to find anomaly. The SVM is generally used as a supervised learning method. In order to decrease misclassified data, a supervised SVM approach with a slack variable is called soft margin SVM. Additionally, single class learning for classifying outliers can be used as an unsupervised SVM. After considering both SVM learning schemes.

4.2 CIDS: A framework for intrusion detection in cloud systems

Each node has two IDSs detectors, CIDS and HIDS. In this way, the node can cooperatively participate in intrusion detection by identifying the local events that could represent security violations and by exchanging its audit data with other nodes. the sharing of information among the following CIDS components: **Cloud nodes**: contains the resources homogeneously accessed through the cloud middleware. **Guest task**: it is a sequence of actions and commands submitted by a user to an instance of VM. **Logs and audit collector**: it acts as a sensor for both CIDS and HIDS detectors and collects logs, audit data, and sequence of user actions and commands. **VM**: it encapsulates the system to be monitored using VMM. The detection mechanisms are implemented outside the VM, i.e. out of reach of intruders. A single instance of a VM monitors can observe several VMs.

4.3 Performance Metric Selection for Autonomic Anomaly Detection on Cloud Computing Systems

To make the anomaly detection tractable and yield high accuracy, the paper apply dimensionality reduction, which transforms the collected health data to a new metric space with only the more relevant attributes preserved. We apply two approaches to reducing dimensionality: metric selection using mutual information and metric extraction by principal component analysis.

4.4 A SVM Model based on Network Traffic Prediction for Detecting Anomalies

The purpose of our Anomaly Detection Mechanism is to provide an efficient method to detect anomalies in the cloud-based network traffic. Figure 1 depicts the basis of our mechanism, by highlighting the application scenario and the main conceptual components.
The cloud provider offers several services by the Internet, such as infrastructure, software and platform to the clients. Real-time cloud traffic data (Flow 1) is continuously being gathered from the cloud environment by the Cloud Monitoring module. This information is subsequently processed by the Poisson-based Predictor that performs prediction based on information such as the protocol type, the number of network packets and timestamp.
After that, the SVM Model is fed with features extracted from the predicted data. Then, the SVM Model triggers a warning to the Event Auditor when an anomalous behaviour is detected. In the meantime, the Repository of Outcomes component stores a detailed output regarding the historic of the Virtual Machine (VM) operation. Furthermore, the Event Auditor represents an agent placed in the VM that is able to communicate collaboratively with agents in the other VMs. This agent receives any anomalous event from the SVM Model and builds a message with information of all components for sending alerts to other agents. Having presented an overview of the anomaly detection mechanism, in the following subsections there will be a more detailed description of the forecasting approach for estimating network traffic on the basis of a Poisson process and the Support Vector Machine model for detecting anomalies in the cloud-based environment.

5 PROBLEM DEFINITION AND SCOPE

5.1 Problem Definition

To design a system to extract the meaningful features from large dataset to increase the efficiency of anomaly detection.

5.2 Scope

The successful attacks causing damages have a high level of effect on the result. Hence to lower down this effect countermeasure play an important role which surpasses the damage done. These countermeasure are responsible for maintaining the effectiveness of results in machine learning

For the above purpose selection of labels from data set is most important task, whole functioning depends on selection of labels. As if wrong features or labels get selected then it will have adverse effect on system performance. Result of anomaly detection will purely depend on how we select the labels to go ahead for other operations.

6 DIFFERENT MACHINE LEARNING ALGORITHM

6.1 Support Vector Machine (SVM)

It is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. It has high prediction accuracy and performance rate but is limited to two classified classes only.

6.2 Decision Tree classifiers

It repetitively divides the working area(plot) into sub part by identifying lines.Operations are carried with optimization.Effciency reduces with increase in dataset.

6.3 KNN

A simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions) Based on optimal solution time complexity is quite high.

7 METHODOLOGY

7.1 Workflow

Figure 1: This is placeholder image: Workflow

7.2 Mathematical model

S = {s, e, X, Y, f_{main}, f_f, DD, NDD, mem_{sh} | ϕ }

s: start state.
e: end state.

Let X be the input set consisting of:- X = L_i
where L is the collected Log from monitoring cloud
Let Y be the output set consisting of:-
Y = C,P where C ∈ Cl is class defined as anomaly.

Functions

f_{main} - Let 'k' be the function to detect the anomaly such that:-

k : log dataset → P

f_f : f_1, f_2

f_1= Cloud Monitoring functions for collecting data

f_2= Anomaly detection function

Success- Failure Rate

P = normal
$P = \phi$
or
P ≠ normal

8 Results

8.1 Data

Table 2: Data Table

S.No	Data set	size
1	KDD1998	43.5 MB
2	KDD 1999	75.3 MB

8.2 Implementation Results

Figure 2: This is placeholder image: Result of KDD 1998 dataset

Figure 3: This is placeholder image: Result of KDD 1999 dataset

9 CONCLUSION

Feature extraction process can affect the system in both ways if the process is not carried out carefully. As features in machine learning is among the important factors which affects the system performance. Feature extraction along with supervised learning algorithm can improve the performance of anomaly detection system to an extent. Reducing the dataset through feature extraction make easy for learning algorithm to focus on important feature and get the work done

References

[1] Dalmazo, Bruno L., et al. "Expedite feature extraction for enhanced cloud anomaly detection." Network Operations and Management Symposium (NOMS), 2016 " *IEEE/IFIP. IEEE, 2016.*

[2] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," Information Sciences, vol. 177, no. 18, pp. 3799 – 3821, 2007. [Online]. Available.

[3] H. Kholidy and F. Baiardi, "CIDS: A framework for intrusion detection in cloud systems," in Ninth International Conference on InformationTechnology: New Generations (ITNG), 2012, April 2012, pp. 379–385.

[4] Fu, Song. "Performance metric selection for autonomic anomaly detection on cloud computing systems." Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE. IEEE, 2011.

[5] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai, and C. D. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines," Expert Systems with Applications, vol. 38, no. 1, pp. 306 – 313, 2011.

[6] P. Ganeshkumar and N. Pandeeswari, "Adaptive neuro-fuzzy-based anomaly detection system in cloud," International Journal of Fuzzy Systems, pp. 1–12, 2015.

[7] B. L. Dalmazo, J. P. Vilela, and M. Curado, "Online traffic prediction in the cloud: A dynamic window approach," in The 2nd International Conference on Future Internet of Things and Cloud (FiCloud'2014), Aug 2014, pp. 9–14.

attach your review and visit log here.....

attach plagiarism report here.....