A SEMINAR REPORT ON

ENHANCING CLOUD ANOMALY DETECTION THROUGH FEATURE EXTRACTION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

MASTER OF ENGINEERING (Computer Engineering)

BY

Nandit Malviya
Exam No. 6907

Under the guidance of
Prof. M.S. Takalikar

DEPARTMENT OF COMPUTER ENGINEERING
Pune Institute of Computer Technology
Dhankawadi, Pune
Maharashtra 411043

DEPARTMENT OF COMPUTER ENGINEERING
Pune Institute of Computer Technology
Dhankawadi, Pune
Maharashtra 411043

CERTIFICATE

This is to certify that the Seminar report entitled
"ENHANCING CLOUD ANOMALY DETECTION THROUGH FEATURE EXTRACTION"

Submitted by
Nandit Malviya Exam No. 8907

is a work carried out by him under the supervision of Prof. M.S. Takalikar and it is submitted towards the partial fulfillment of the requirement of Savitribai Phule Pune University, Pune for the award of the degree of Master of Engineering (Computer Engineering)

Prof. M.S. Takalikar
Internal Guide
PICT, Pune

Dr. R. B. Ingle
Head
Department of Computer Engineering
PICT, Pune

Place:
Date:

ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Prof. S. S. Sonawane and Head of Department Dr. R. B. Ingle for their support.

I also sincerely convey my gratitude to my guide Prof. M.S. Takalikar, 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.

Above all I would like to thank my parents for their wonderful support and blessings, without which I would not have been able to accomplish my goal.

Contents

List of Tables

List of Figures

Abstract

In Computer Engineering Machine learning is being used in a wide range of application domains to discover patterns in large datasets. Machine learning consolidation with cloud computing is increasing day by day. With increase of data over cloud and managing of data has involved other technology like machine learning, deep learning to increase security of cloud data.

Cloud computing is the latest trend in business for providing software, platforms and services over the Internet. However, a widespread adoption of this paradigm has been hampered by the lack of security mechanisms. In view of this, the aim of this work is to propose a new approach for detecting anomalies in cloud network traffic. The anomaly detection mechanism works on the basis of a Support Vector Machine (SVM). The key requirement for improving the accuracy of the SVM model, in the context of cloud, is to reduce the total amount of data.

The labelled data set has label that help in extracting feature and decrease overall data set.For data collection cloud has to be monitored. For feature extraction Poison Moving Average is being used.

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:

No.	Techniques	Cloud scenario	Feature extracted	dataset	limitations
1	Poisoning Moving Average, Support vector machine	yes	Protocol type, port number, packet size, number of packets	DARPA, CMU	classifies data into 2 classes only
4	Decision tree classifier, Maximal relevance and minimal redundancy	yes	CPU usage, memory ,swap utilization, paging and paging faults	Own Dataset, KDD	It uses its own dataset so specialised data accuracy is not there
3	Event correlatio. ANN	yes	User logs and signature	Own Dataset, KDD	Features extracted are on the basis of own dataset used.
2	Genetic algorithm, SVM	No	Protocol, source port, destination port, IP, TTL	DARPA	Cloud deployment is not done.
5	Hierarchical clustering Algorithm, SVM	No	Protocol, duration of connection, status	DARPA	cloud deployment not present and efficiency is less

Table 1: Literature survey

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 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

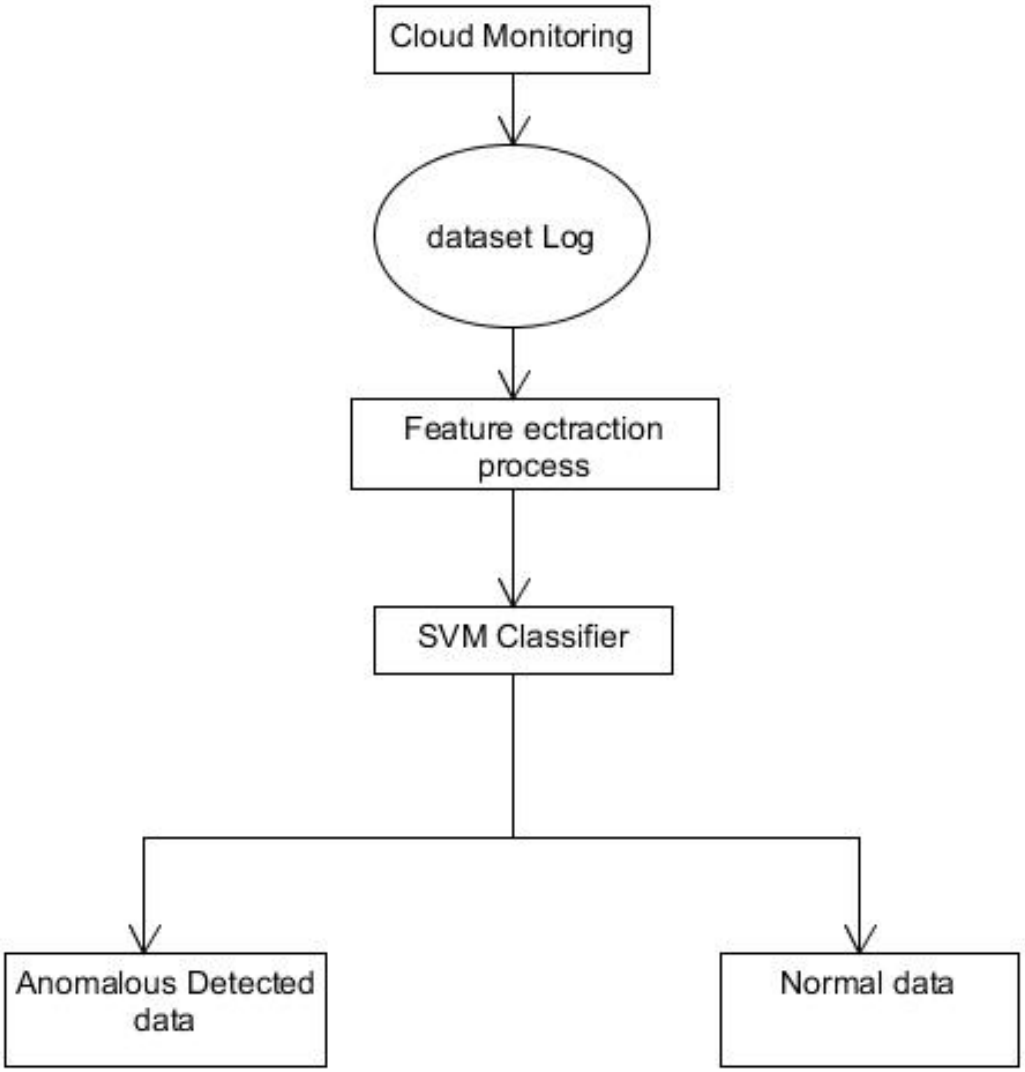

Figure 1: Workflow

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 = ϕ
or
#$P \neq$ normal

8 Results

8.1 Data

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

Table 2: Data Table

8.2 Implementation Results

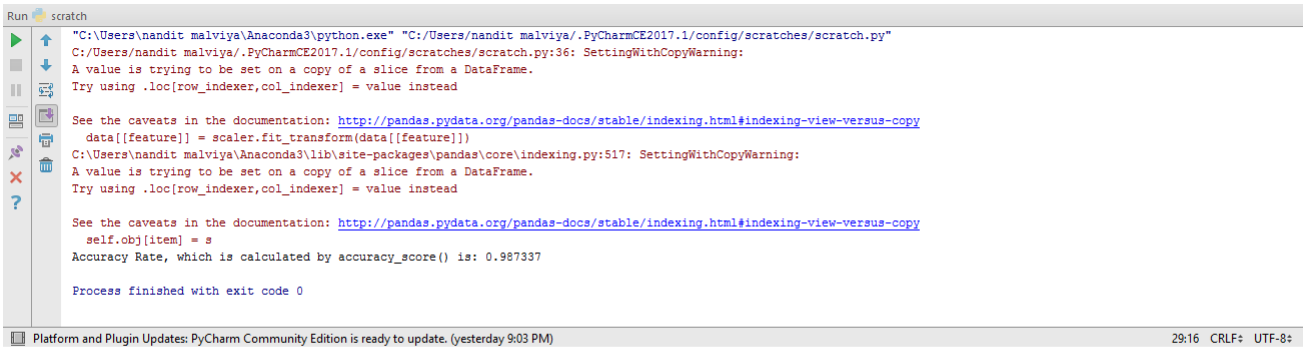

Figure 2: Result of KDD 1998 dataset

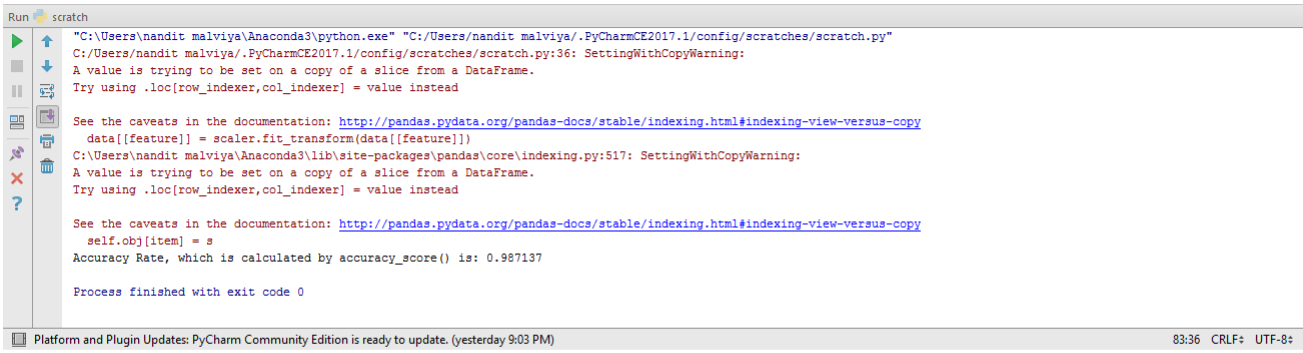

Figure 3: Result of KDD 1999 dataset

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.