The False Positive Alert Reduction Using Data mining Techniques in Intrusion Detection System

Anthony Raj. A
Research Scholar / CA
PRIST UNIVERSITY, Thanjavur,
Asst. Professor Dept of CS, SBMJC, KGF

Siddarama. S
Asst. Professor & Head, Dept of CS,
Sri Bhagawan Mahaveer Jain College
KGF, Karnataka., INDIA

Abstract- Information security is a vital aspect of any organization. Most of the organizations relay and trust on the intrusion Detection System (IDS) which play important role in detecting intrusions in data network environment. The design of IDS varies with implementation of different IDS techniques involved. The design of IDS techniques keep changes as the trend of data network innovative attack methods gets updated day by day. Hence there is no single perfect solution is found for detecting the intrusions in the data network. In general IDS systems are complex and it is an ongoing process. There are dissimilar types of intrusion detection systems exist and pass through a common problem of rendering high volume of alerts and immense number of false positives. The false positive alert alters the space and time complexities of the IDS modules and gradually slows down the detection rate and performance of the system. This is the main motive behind the research of this paper. The objective of this research paper is to explore and suggest different techniques which help design in building the optimal intrusion Detection system of low cost and high performing computational capability and adaptability to various network environments for the results of false alert reduction, a high intrusion detection rate, risk management both detection and control the intrusions, finally to identify the real attacks from other false alarms and events of the system. This paper navigates through different associated studies of the last decade with providing a citation for further research in this domain. Various unresolved issues have also been covered in this manuscript.

Keywords: alert reduction, clustering, alert correlation, adaptive learning, post processing.

1 INTRODUCTION

In the beginning Anderson presented IDS system and later it is validated by Dennis. The most important trouble faced by the IDS is false positive that is misidentified, considered as security issues and this draws the more interest from the intrusion detection analyst. The calculations reveal that 99% of alerts that are reported by IDS are not associated with security issues. The main reasons for this found to be the following factors: Specificity of detection signatures, Runtime limitations, Base-rate fallacy, Dependency on environment. [1]

From the above information it obviously difficult to build the IDS which have very small number of positives and is highly unmanageable job. The research paper represents the recap of the current research methods and techniques that are related to the false positives reduction problem and reviews the basic idea of using existing IDSs as an alert source for applying complementary approaches like off-line using data mining techniques with online using machine learning alert processing techniques to bring down the number of false positives. [1]

2. RELATED WORK

The related work in alert management includes improving the quality of alerts and alert correlation. In the first category that try to improve the caliber of alerts by supplying additional information, such as alert vulnerability statements and context. Such an idea has been applied by Paxson and Sommer to improve Bro's byte-level alert signatures with context derived from protocol analysis and regular expressions. Another approach is to match alerts with exposure reports carried out by commercial alert correlation products. Lippmann et al. indicate how such data can be utilized to prioritize alerts, depending on the exposures of the target: Even correctly identified signs of intrusions can be given a lower priority or even be thrown away if the target has been identified not to be vulnerable to that specific attack. A conventional model for alert correlation and supplementing such information has been presented by Morin et al. [1, 3]

The second approach is obtained through intuition rather than from reasoning or observation, and proved to be good in real environments. As presented by Wespi and Debar given three alert dimensions (namely destination address, source address and attack class), alerts can be sorted into settings depending on the number of matching dimensions from the near general (only one dimension matches) to the most particular case (all three dimensions match). Each of these settings has an actual meaning. This approach is successfully applied in alert pre-processing, in spite of the fact that it can be limited when correlating to a greater extent complex attacks (e.g., an attempt compromising the host and then setting up another attempt from there).[1]

Investigators have discovered that preprocessing is demanded for better answers and practiced several approaches. Late Intrusion Detection Systems can flood out with amount of information they ought to study. This difficulty is considered by Fernando. They have talked about that we call for to eliminate unauthentic and surplus information from new data before applying it for intrusion detection. They use user Rough Set for key attribute recognition. Using n-gram theory they have recognized surplus subsequences. They have also presented Hidden Markov Model for service selection. Using observational answers they indicate how their approach brings down examine rate significantly. [1, 3]

A new cooperating filtering technique for preprocessing the investigation type of attacks is presented by G. Sunil Kumar They carried out a hybrid classifiers based on binary particle stream optimization and random forests algorithm for the sorting of examine attacks in a network. They used worldwide search potentiality of particle stream optimization when random forests utilized as a classifier. Their observational result proved that as number of trees utilized in forest step-ups, the false positive rate step-downs.

Preprocessing of web server log file is presented by Shaimaa for better quality of data and accordingly better mining result. They have blended dissimilar web log files with dissimilar formats in one integrated format using XML. It will help in going after drawing out more plans of attack. Shaimaa has talked about ambitious data of web log file and noisy data hinted preprocessing to get rid of such condition of being impure. Priyanka et al has also believed almost similar thought and suggested web log preprocessing for caliber results. [3]

Salem et al indicated preprocessing approximate network traffic in to association records. Applied tool can furnish relevant and summarized information for intrusion detection. Zheng has advised Hierarchical Intrusion Detection. It applies neural network and statistical preprocessing classification. They have examined dissimilar types of neural network classifiers and also did stress test. Sanjay offered Singular Value splitting as a preprocessing step to bring down the spatial property of data. Such step-down will contribute importance to more outstanding features in data [1,3]

Data mining techniques for the first time applied for cognition discovery from telecommunication event logs more than a decade ago. In the setting of IDS alerts mining, a number of techniques have been proposed. Clifton and Gengo have looked into the detection of frequent alert successions and enhanced by Ferenc, Walter A. Kosters and Wim Pijls, used this cognition for making IDS alert filters. Long et al. indicated a clustering algorithm for differentiating Snort IDS true alerts from existing false positive alerts. Julisch and Dacier also projected a abstract clustering technique for IDS alert logs, thus clusters match to alert descriptions, then human experience can be applied for formulating correlation and filtering rules for next IDS alerts. Many of other advances have been hinted like time series modeling, machine learning, and the use of the control charts, etc.[3,8]

3. UNIFIED APPROACH FOR REDUCTION OF FALSE POSITIVE ALERT

We introduce reviewed bi-complementary paths to bring down the count of false positives in intrusion detection applying alert post processing through machine learning and data mining. Furthermore, these both techniques, because of their complementary type, can be employed together in an alert-direction system. These constructs have been maintained on a variety of data sets, and accomplished a significant step-down of the number of false positives in both simulated and real surroundings. [2, 1]

3.1 Surveyed Data-Mining Approach - Clustering Alerts for Root Cause Analysis(CLARAty)

The end of the CLARAty (Clustering Alerts for Root Cause investigation) is to efficiently discover large clusters of alerts and to report them in a derived way in order to build commonalities among the dissimilar alerts expressed and apprehensible for a human. Conforming to an ultimate standard, these derived reporting's represent to root causes and are in act actionable for the IDS security expert who is skilled at analyzing data.

Figure 1 depicts how CLARAty matches into the ideal alert-direction establish. Its constituents and work flow: historic alerts, which are immediately yielded by the IDSs in the reporting surroundings and are hence untagged, are exploited for large clusters of alerts (check below for inside information). The reporting's of these clusters are delivered in the form of derived alerts to a data analyst to look for the fundamental reasons. If these reasons are regarded to be false positives instead signs of real attacks, the alerts should not load by the operators any longer in the future and should therefore be moved out.[3,1]

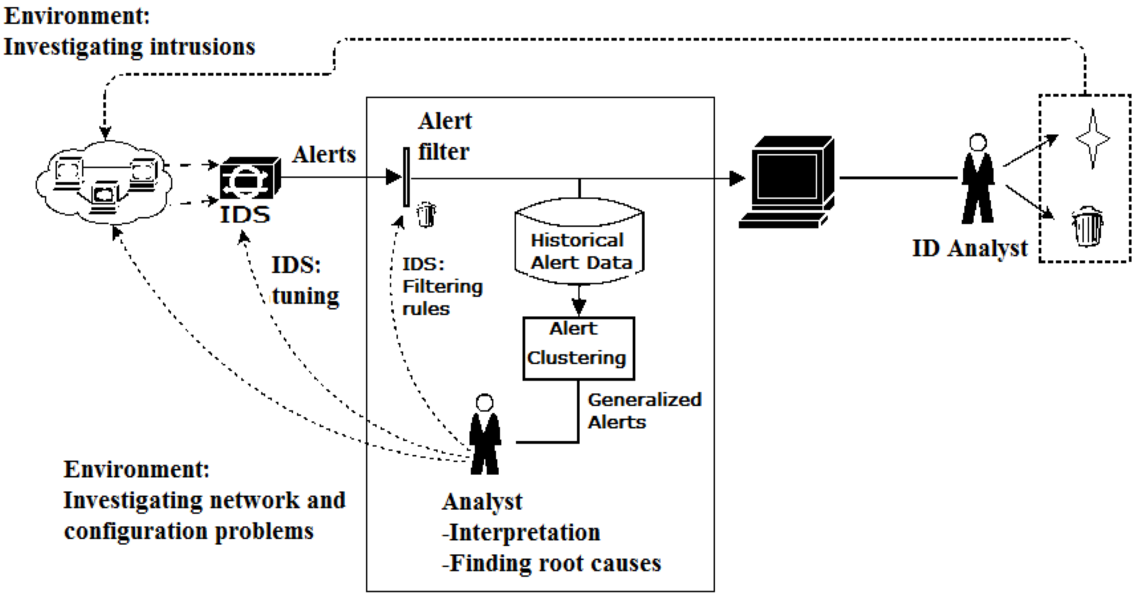

Figure 1. Alert direction system using CLARAty (data mining) to expose reasons and bring down false positives in intrusion detection

Numerous dissimilar data-mining techniques make it for cluster analysis. The suitableness of the dissimilar methods powerfully looks on the area of covering and its dimensions. For our job of alert clustering, where we expect for human apprehensible Sort of alert clusters inspecting to main reasons, attribute-oriented induction (AOI) with annexes for this particular practical application area is the most conditioning. The changed AOI algorithm as described applies the abstraction hierarchies reporting the circumstance knowledge to merge alerts into inferred alerts iteratively. These derived alerts contain at least partially inferred attributes, i.e., attributes derived through the above hierarchies outside the lowest level.

As an example, check the first derived alert in Table 1: It reports a cluster of alerts of the type WWW IIS view source attack coming from a Non Privileged port of a machine in the Internet, pointing the particular destination address ip5 on the particular port 80. In this case the timestamp was extrapolated to the highest level anytime, pointing that the alerts happen at random clocks and no preference for particular days of the week or times throughout the day could be detected.[6,1]

TABLE 1 EXAMPLE OF GENERALIZED ALERTS IN THE INTRUSION DETECTION LOGS

ALERT TYPE	SOURCE PORT	SOURCE IP	DEST PORT	DEST IP	TIME	SIZE (TOTAL 156380)
WWW IIS view source attack	Non Privileged	Internet	80	ip5	Anytime	54310
FTP SYST command attempt	Non Privileged	Firewall	21	Internet	Anytime	6439
IP fragment attack	n/a	ip6	n/a	Firewall	Workday	4581
TCP SYN host sweep	Non Privileged	Firewall	25	AnyIP	Anytime	761

3.2 Machine Learning approach-Labeled alert

The second plan of attack covers the problem of false positives in intrusion detection by constructing an alert classifier that distinguishes true from false positives. We specify alert classification as binding a label from a defined set of user-defined labels to an alert. The alerts are separated into false and true positives, but the sorting can be continued to point the class of an attack, the cases of a false positive or an indefinite thing.

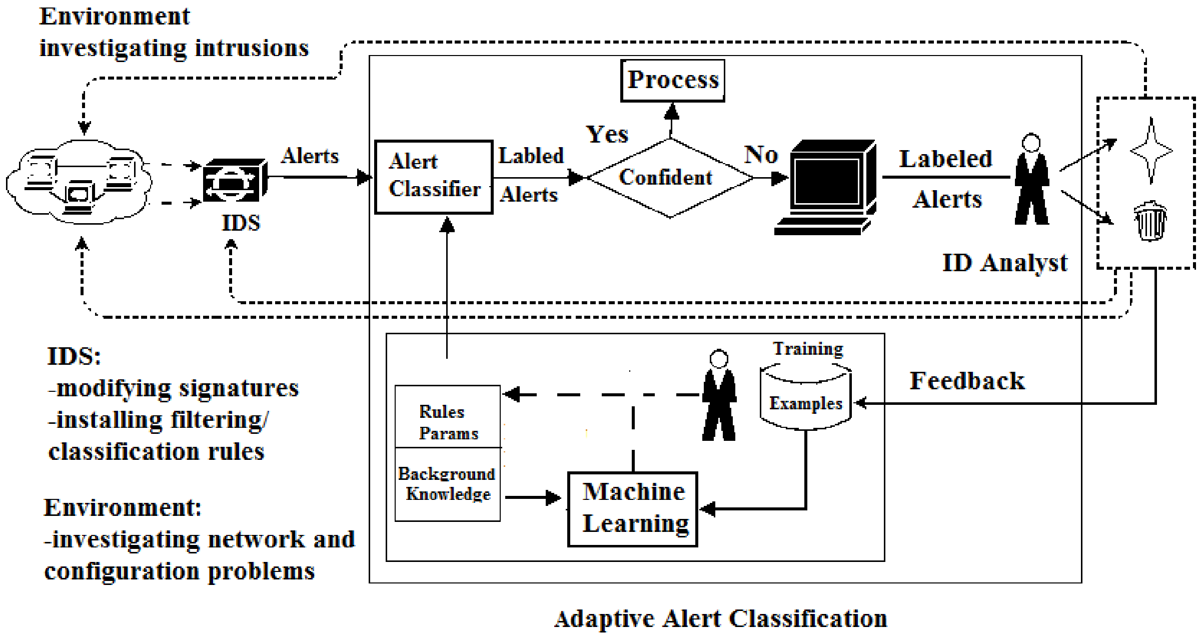

Figure 2. Alert management system using ALAC (Machine learning) to built an adaptive alert classifier

Alerts are sorted out by a so-called alert classifier. Alert classifiers can be developed mechanically using machine-learning techniques or they can be developed manually by human souls. The Adaptive Learner for Alert Classification (ALAC) inserted in functions the previous approach. Most significantly, ALAC discovers alert classifiers having precisely and clearly conveyed classification logic so that a human expert can audit it and confirm its rightness. Likewise, the analyst can acquire confidence in ALAC by realizing how it works.

ALAC classes alerts into true positives and false positives, and demonstrates these classes to the intrusion detection analyst, as depicted in Fig. 2. In dividing line to the standard alert management approach, ALAC applies the feedback of the analyst, who is classing the alerts at the alert console, to produce labeled alerts. These labeled alerts are utilized by the system to give training examples, which are employed by machine learning techniques to first construct and then modify the classifier. The classifier is then employed to classify new alerts. This process is continuously iterated to amend the alert classes. The analyst can go through the classifier at any time.[4,3,1]

The machine-learning techniques mainly employed to construct an alert classifier system which able to distinguishes true out of false positives. Alert classification system does pose various issues. The spot of classes are commonly much oblique, i.e., false positives are more common than true positives. The cost of misconstruction of alerts are (i.e., misclassifying true positives as false positives) usually more high-priced than vice versa. ALAC sorts out alerts in real-time and modifies its classifier as new alerts become available.

The learning technique should be effective enough to function in real time and incrementally, i.e., to be able to change its logical thinking as new data gets available. Fourth, we expect the machine learning technique to apply background knowledge, i.e., additional information such as alert database, network topology, alert context, etc., which is not checked in alerts, but admits us to construct more exact classifiers (e.g., classifiers using derived concepts, alike to those employed in CLARAty). As a matter of fact, search in machine learning has depicted that utilize of background knowledge oftentimes contributes to more instinctive and brief rules. Nevertheless, utilize of background knowledge gains the complexness of a discovering task, and only some machine-learning techniques bear it.[1,6]

In machine discovering, if the discoverer has no prior knowledge about the learning job, it finds only from study cases. nevertheless, hard learning jobs typically call for a essential body of prior information, which causes it possible to convey the learned construct in a more natural and brief manner. In the field of machine learning such data is cited to as background data, On the other hand in the field of intrusion detection it is quite often named as circumstance information. The utilize of background data is also very significant in intrusion detection. Examples of background data include: Alert Context, Network topology, installed software and Alert Semantics.[1]

4. FINDINGS OF THE RESEARCH STUDY

4.1. Data mining and alert correlation techniques

The study aimed to reduce false positives using integrated data mining and alert correlation techniques. We analyzed integrated general approaches namely the detection techniques that act during detection phase and the alert processing techniques that are applied on generated alerts after detection phase. The above aspects of existing alert correlation and data mining technique to overcome the IDS‘s problems are discussed. The other aspects of some open problems and disadvantages related to the studied techniques are identified and first; most of the proposed techniques act in an off-line mode. Second, some of these techniques are depended to human analyst for training phase or developing filtering rules and some of the proposed techniques also lack accuracy. To get over the above limitations a new approach of integration suggested by using a SIEM system which merges data mining and alert correlation techniques to improve the IDS system to reduce the false positive rate.[4, 5]

4.4. Identification of reduction spot

The research had been done in the false alert reduction in IDS area. The study suggests the approaches for minimizing the false positives. The suggested technique also considers the attack which is yielded employing a spoofed IP address. The false positive step-down can be discovered in the sensor level or later on at detection level, while at the sensor level can be studied as improving the detection method. So we think that normalization is needed to elucidate false positive reduction condition. In the end the IDS research scholars still keep inventing to find out the most suitable method to decrease the false positive alert and reaction of attacks so that they can be able to block and prohibit these attacks to arrive at the last stage. [2]

4.3. Data preprocessing

We have talked about preprocessing module for minimizing false positive rate in Intrusion Detection System. The research study reviewed the four major functionalities as part of preprocessing module. Sacking of noise and incomplete information, configuration parameter based processing, a prominent attribute selection and extraction, and interconnected log suggested for decrease of false positive rate. The empirical results have indicated that such preprocessing efforts can serve Intrusion Detection System in decreasing the false positives. By consuming more resources and employing efficient complex algorithms for preprocessing in future will further help in reducing false positive. [3]

4.5. Alert post processing

We established the universal picture of alert direction in intrusion detection and introduced two complementary and orthogonal approaches to bring down the number of false positives in intrusion detection: CLARAty, established on root-cause discovery and data mining, and ALAC, established on machine learning. We also proved how unified techniques match into standard alert-management systems.

CLARAty is alert-clustering plan using data mining with a altered version of attribute oriented induction. Applying background information in the form of abstraction hierarchies for the alert properties, it examines historic alert logs for big clusters of alerts expressible by a inferred alert which a human expert can understand to discover root causes. Experimentations with real life data sets have proven that already few lots of inferred alerts cover up over 90% of the new alerts. These inferred alerts can so be realized as root causes, and thus be applied, by filtering or fixing of these root causes, for later alert reduction (accomplishing on mean a 75% reduction filtering out every month).[2,1]

An adaptive alert classifier ALAC is constituted on the resubmit of an intrusion detection analyst and machine-learning techniques. We talked about the vastness of background cognition and why the categorization of IDS alerts is a unmanageable machine-learning problem. At last, we introduced a prototype execution of ALAC and assessed its functioning on a synthetic and a real life intrusion data set. ALAC can function in two ways: a advocate mode, in which all alerts are marked and passed onto the analyst, and federal agent mode, in which some alerts are refined mechanically.

We expressed that the system is useful in advocate mode, where it accommodated to learn the categorization from the analyst. In this fashion we found false positive and false negative rates matching to those of batch categorization. We also discovered that our system is useful in the agent style, where approximately alerts are autonomously refined (e.g., false positives categorized with high assurance are thrown-away). More significantly, for both the false negative rate, data sets of our system are able to compare to that in the advocate mode. At the same instance, the volume of alerts for the analyst to manage has been decreased by 50% and more. We also talked about how both ALAC and CLARAty can be treated in a two-staged alert filtering and categorization system. Hereafter we are designing to measure the functioning of such a system to interpret the possible fundamental interaction and synergies i.e. to produce an effect greater than the sum of their individual effects. We are also mindful of the limits of the data sets applied with ALAC. We direct to assess the execution of the system on the ground of more naturalistic intrusion spying data and to incorporate an alert-correlation system to bring down superfluousness in alerts. We are also expecting at the background information and how it can be better constituted for machine-learning algorithms. [1,2]

5. CONCLUSION AND FUTURE WORK

We put forward in the introduction, ALAC and CLARAty are two negation of the other and can be employed united in a two-staged alert separating out and sorting system. The system would apply CLARAty in the initial stage to regularly extract raw alerts, detect their reasons and either takes out them or establish alert filters. The end product from the first stage would then be sent on to ALAC acting with an operator. The gain of this method is that alert filters out from CLARAty take out the most dominant and wearisome false positives, which effectively amends class dispersion in favor of true positives in the alerts communicated on to the second stage. Furthermore, ALAC obtains fewer alerts to work on, that is significant as of runtime necessities. In the current prototype, we only measured ALAC and CLARAty individually, as the data sets applied for measure of CLARAty were not tagged and thus could not be utilized with ALAC. Then again, the data sets applied for the valuation of ALAC did not hold a adequate number of alerts to vary an efficient data mining with CLARAty. An assessment of the two stages system is imparted as future work.

ACKNOWLEDGMENTS

First authors give Praise and Glory to the Lord Almighty for He is forever good!!! The authors would wish to give thanks Prof. Tadeusz Pietraszek and Axel Tanner IBM Zurich Research Laboratory,Saumerstrasse 4, 8803 Ruschlikon, Switzerland for their worthful study and reappraisals on this domain of work which forms the foundation for this research paper entitled. Finally I thank Prof. Dr.Arul Sir who is my lead and other unidentified authors for their thoughts, boost and supporting.

REFERENCES

[1]. Tadeusz Pietraszek and Axel Tanner IBM Zurich Research Laboratory, Saumerstrasse 4, 8803 Ruschlikon, Switzerland " Data Mining and Machine Learning Towards Reducing False Positives in Intrusion Detection" Appearing in Information Security Technical Report, 10(3), 2005

[2]. Manish Kumar, Dr. M. Hanumanthappa, Dr. T. V. Suresh Kumar 1Asst.Professor, Dept. of Master of Computer Applications, M. S. Ramaiah Institute of Technology, Bangalore-560 054, INDIA" Intrusion Detection System - False Positive Alert Reduction Technique" in ACEEE Int. J. on Network Security , Vol. 02, No. 03, July 2011

[3]. Dharmendra G. Bhatti Associate Professor, Shrimad Rajchandra Institute of Management and Computer Application, Bardoli, Gujarat, India." Data Preprocessing for Reducing False Positive Rate in Intrusion Detection" in International Journal of Computer Applications (0975 – 8887) Volume 57– No.5, November 2012.

[4]. Asieh Mokarian, Ahmad Faraahi, Arash Ghorbannia Delavar, Payame Noor University, Tehran, IRAN " False Positives Reduction Techniques in Intrusion Detection Systems-A Review" in IJCSNS International Journal of Computer Science and Network Security, VOL.13 No.10, October 2013

[5]. EMostapha Chakir , Chancerel Codjovi, Youness Idrissi Khamlichi, Mohammed Moughit at Laboratory of Computer Networks, Mobility and Modeling Faculty of Science and Technology University Hassan First Settat, Morroco "False Positives Reduction in Intrusion Detection Systems Using Alert Correlation and Data mining Techniques" in International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 4, 2015 ISSN: 2277 128X

[6]. Mahdi Zamani and Mahnush Movahedi {zamani,movahedi}@cs.unm.edu Department of Computer Science University of New Mexico "Machine Learning Techniques for Intrusion Detection" in arXiv:1312.2177v2 [cs.CR] 9 May 2015

[7]. Salma Elhag, Alberto Fernández , Abdullah Bawakid, Saleh Alshomrani , Francisco Herrera c,d aDepartment of Information Systems, King Abdulaziz University (KAU), Jeddah, Saudi Arabia Department of Computer Science, University of Jaén, Jaén, Spain "On the combination of genetic fuzzy systems and pair wise learning for improving detection rates on Intrusion Detection Systems" in Expert Systems with Applications journal homepage: www.elsevier.com/locate/eswa. Expert Systems with Applications 42 (2015) 193–202

[8]. Hany Nashat Gabra Computer and Systems Engineering Department, Ain Shams University, Cairo, Egypt. hanynashat@hotmail.com Dr. Ayman M. Bahaa-Eldin Computer and Systems Engineering Department, Ain Shams University, Cairo, Egypt. ayman.bahaa@eng.asu.edu.eg Prof.Huda Korashy Computer and Systems Engineering Department, Ain Shams University, Cairo, Egypt. hoda.korashy@eng.asu.edu.eg "Classification of IDS Alerts with Data Mining Techniques"