INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES

ISSN: 2322-0821(O) VOLUME 5 ISSUE 2
ISSN: 2394-9910(P) Apr 2017 - Jun 2017

www.irjeas.org

IMPROVED DECISION TREE CLASSIFICATION PERFORMANCE

Babita Patel[1] Anurag Shrivastava[2], Jyoti Sondhi[3]

[1]*M.Tech Scholar, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[2]*Asso. Professor &Head, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[3]*Assistant Professor, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*

***Abstract*—Growth and popularization of information system increase now days, security of information is a big problems. Intrusion Detection System (IDS) as the main security defensive technique and is widely used against intrusion. Data Mining and Machine Learning techniques proved useful and attracted increasing attention in the network intrusion detection research area. Recently, many machine learning methods have also been applied by researchers, to obtain high uncovering rate and low false alarm rate on KDD CUP'99 dataset used for intrusion detection system. Unfortunately a potential drawback of all those methods is that how to classify attack or intrusion effectively. Use of internet is increasing progressively, so that large amount of data and it security is also an issue. Another problem with KDD Cup 99 Dataset is class imbalanced. Sampling technique is one the solution of large dataset and class imbalanced. This work proposes a sampling technique for obtaining the sampled data. Sampled dataset represent the whole dataset with proper class balancing. Imbalanced classes can be balanced by sampling techniques. The purpose of this paper is to propose IDS framework model based on proposed sampling, class balancing and machine learning technique. This model improves the classification performance. The Proposed work is tested on basis of Accuracy, Error rate, Detection rate and False Alarm rate.**

***Keywords*— Class Balancing, Sampling, Classification, Machine learning technique, IDS.**

I. Introduction

We securing information either in private or government sector has become an essential requirement. System vulnerabilities and valuable information magnetize most attackers' attention. Traditional intrusion detection approaches such as firewalls or encryption are not sufficient to prevent system from all attack types. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security layer against these malicious or suspicious and abnormal activities. Thus, intrusion detection system (IDS) has been introduced as a security technique to detect various attacks. IDS can be identified by two techniques, namely misuse detection and anomaly detection. Misuse detection techniques can detect known attacks by examining attack patterns, much like virus detection by an antivirus application. However they cannot detect unknown attacks and need to update their attack pattern signature whenever there is new attacks .On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage as intrusion. Although anomaly detection has the capability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate .In recent years, and interest was given into machine learning techniques to overcome the constraint of traditional intrusion techniques by increasing accuracy and detection rates. New machine learning based IDS with sampling is used in our detection approach. The advantage of IDS (Intrusion Detection system) can greatly reduce the time for system administrators/users to analyze large data and protect the system from illicit attacks. Improve the performance of IDS and the low false alarm rate.

A. Data Mining

Data Mining is defined as the technique of extracting information or knowledge from huge amount of data. In other words, we can say that data mining is mining knowledge from large data.

B. Machine Learning Technique :

When a computer needs to perform a certain task, a programmer's solution is to write a computer program that performs the task. A computer program is a piece of code that instructs the computer which actions to take in order to perform the task. The field of machine learning is concerned with the higher-level question of how to construct computer programs that automatically learn with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E Thus, machine learning algorithms automatically extract knowledge from machine readable information. In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents (task T). A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. The training experiences (E) could be labeled instances.

II. Related work:

The authors [1] have proposed to use data mining technique including classification tree and support vector machines for intrusion detection. Utilize data mining for solving the problem of intrusion because of following reasons: It can process large amount of data. User's subjective evolution is not necessary, and it is more suitable to discover the ignored and unknown information. Machine learning based ID3 and C4.5 two common classification tree algorithms used in data mining. Author said C4.5 algorithm is better than SVM in detecting network intrusions and false alarm rate in KDD CUP 99 dataset.

In [2], the author said performance of a Machine Learning algorithm called Decision Tree is evaluated and compared with two other Machine Learning algorithms namely Neural Network and Support Vector Machines which has been conducted by A. The algorithms were tested based on accuracy, detection rate, false alarm rate and accuracy of four categories of attacks. From the experiments conducted, it was found that the Decision tree algorithm outperformed the other two algorithms. Compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

In [3], the authors have proposed supervised learning with pre-processing step for intrusion detection. Authors used the stratified weighted sampling techniques to generate the samples from original dataset. These sampled applied on the proposed algorithm, proposed method used the stratified sampling and decision tree. The accuracy of proposed model is compared with existing results in order to verify the validity and accuracy of the proposed model. The results showed that the proposed approach gives better and robust representation of data. The experiments and evaluations of the proposed intrusion detection system are performed with the KDD Cup 99 dataset. The experimental results show that the proposed system achieved higher Accuracy and Low Error in identifying whether the records are normal or attack one.

In [4] authors said today's era data and information security is most important. The Intrusion detection system deals with large amount of data which contains various irrelevant and redundant features resulting in increased processing time and low detection rate. Therefore feature selection plays an important role in intrusion detection. There is various feature selection methods used. Author's comparer the different feature selection methods are presented on KDDCUP'99 dataset and their performance are evaluated in terms of detection rate. Out of the total 41 network traffic features, used in detecting intrusion, some features will be potential in detecting intrusions. Therefore the predominant features are extracted from the 41 features that are really effective in detecting intrusions. Feature selection can reduce the computation time and model complexity.

In [5] authors said Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives which indicates the complete data set by examining a fraction. This paper focuses on different types of sampling strategies applied on neural network. Here sampling technique has been applied on two real, integers and categorical dataset such as yeast and hepatitis data set prior to classification. Authors give the comparison of different sampling strategies for classification which gives more accuracy.

In [6] authors investigate the effect of sampling methods on the performance of quantitative bankruptcy prediction models on real highly imbalanced dataset. Seven sampling methods and five quantitative models are tested on two real highly imbalanced datasets. A comparison of model performance tested on random paired sample set and real imbalanced sample set is also conducted. The commonly used re-sampling strategies include oversampling and under-sampling. Two widely used oversampling methods: Random Oversampling with Replication (ROWR) and Synthetic Minority Over-sampling Technique (SMOTE) are employed in this paper. Two Under-sampling sampling method: Random under sampling (RU) and Under-sampling Based on Clustering from Gaussian Mixture Distribution (UBOCFGMD). Under-sampling method is better than oversampling method because there is no significant difference on performance but oversampling method consumes more computational time.

The work [7] discusses imbalanced dataset. A dataset is imbalanced if the classification categories are not approximately equally represented. Authors discuss some of the sampling techniques used for balancing the datasets, and the performance measures more appropriate for mining imbalanced datasets. Over and under-sampling methodologies have received significant attention to counter the effect of imbalanced data sets. Sampling methods are very popular in balancing the class distribution before learning a classifier.

III. DATA SET AND SAMPLING:

A. KDD CUP 99 DATASET: Used in the evaluate machine learning technique. In practice, we recognize that this dataset is decade old and has many criticisms for Current research. But we believe that it is still sufficient for our experiment which aims to reflect the performance of distinct machine learning approaches in a general way and find out relevant issues. In addition, the full KDD99 dataset Contain 4,898,431 records and each record contain 41 features. Due to the computing power, we do not use the full dataset of KDD99 in the experiment but a 10% portion use of it. This 10% KDD99 dataset contains 494,021 records (each with 41 features) and 4 categories of attacks. The details of attack categories and specific types are shown in Table1. According to Table1, there are four attack categories in 10% KDD99 dataset:

(1) Probing: Scan networks to gather deeper information
(2) DoS: Denial of service
(3) U2R: Illegal access to gain super user privileges
(4) R2L: Illegal access from a remote machine.

The number of samples of various types in the training set and the test set are listed respectively in tables below:

NORMAL	Attack				Total
	DoS	U2R	R2L	PROBE	
	391458	52	1126	4107	
97278	396743				494021

Table 1.1 Dataset Descriptions

B. Sampling:

Data sets contain very large amount of data which is not an easy task for the user to scan the entire data set. The researcher's initial task is to formulate a rational justification for the use of sampling in his research. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. It is the process of selecting representatives which indicates the complete data set by examining a fraction. Due to sampling we overcome the problems like; i) in research it is not possible to collect and test each and every element from the data base individually; and ii) study of sample rather than the entire dataset is also sometimes likely to produce more reliable results.

C. Class Imbalanced:

A Dataset is imbalanced if the Classification categories are not just about equally represented. Over and under-sampling methodologies have received attention to counter the effect of imbalanced data sets[10].

D. Feature selection

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s [9]. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features [10] from the data for building an effective and efficient learning model [11].

A number of feature selection algorithms are proposed by various authors.[] Attribute evaluator is basically used for ranking all the features according to some metric.

IV. PROPOSED WORK

Some research in machine learning community has addressed the strategy of re-sampling the original dataset to deal with the issue of class imbalance []. The commonly used re-sampling strategies include oversampling and under-sampling. Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes, while under-sampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In the original imbalanced training dataset, let the original sample set of minority class and majority class denoted by Cmin and Cmax separately, the size of minority class Cmin is much less than the size of majority class Cmax.

In the KDD cup 99 data set DoS is a majority class and U2R and R2L is the minority class. Two other classes' normal and probe assume as the optimal and other classes. Therefore, the set of minority class Cmin = {I1, I2} and Cmin = 2, the set of majority class Cmax = {M1} and Cmax = 1. Copt is the set of optimal class Copt =1. Cother is another class.

Under sampling:

Under sampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In Random under sampling the majority class is under-sampled by randomly removing samples from the majority class Population until the majority class becomes minority class.

Over sampling:

Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes.

Proposed Algorithms:

1. *Let C1, C2 …Cn are classes.*
2. $\sum_{p=1}^{n} CEp = Dataset\ Element,\ where\ p = 1,2,3 \ldots n.$
3. *Select and make classes with majority, Minority, optimal & others.*
4. *Let Cmax Represents Majority*

   Cmajority = (Cmax)

   Cminority = (Cmin)

   Coptimal = (Copt)
5. *Under-sample the majority (Cmax) up to optimal (Copt)*
6. *Over-sample the minority (Cmin) up to optimal (Copt)*
7. *Get sampled dataset with balanced classes*
8. *Now applying feature selection method*

In describing our experiments, our terminology will be such that if we under-sample the majority class up to optimal class and over-sampled the minority class up to optimal class. By applying a combination of under-sampling and over-sampling, the initial bias of the learner towards the negative (majority) class is reversed in the favor of the positive (minority) class.

V. ARCHITECTURE OF THE PROPOSED MODEL

In Architecture of the Proposed model shows that in 10% portion of KDD99 dataset Firstly we are applying sampling and class balancing technique and get balanced sampled dataset now we are using preprocessing technique in sampled and balanced dataset and applying feature selection method.

Now going to classification part and determine the training and testing data in very short period after that applying classification technique in trained data and evaluate the result. Same procedure is applying in different machine learning classifier and measure result. Also measure the classifier performance with un-sampled and imbalanced dataset.

Parameter of the performance measures in the terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

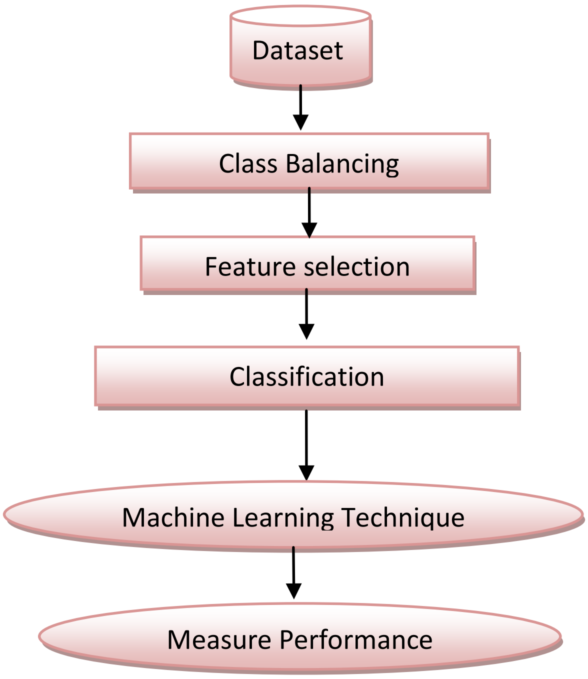

Figure 1. Architecture of the system

VI. RESULT ANALYSIS

Balanced sampled KDD'99 dataset, obtain from proposed sampling technique give better result. Table 6.1 show the actual result of decision tree classifier on sampled and un-sampled dataset.

Parameter	Decision Tree (Sampled)	Decision Tree (UnSampled)
Accuracy	99.60%	98.36%
Error Rate	0.39%	1.63%
Detection Rate	82.12%	81.24%
False alarm Rate	0.39%	1.09%

Table 6.1 Result of Decision Tree

Result shows the performance of the proposed approach classifier in terms of accuracy and error rate on sampled and un-sampled KDD, 99 dataset. Result also shows performance of the different machine learning classifier in terms of the detection rate and false alarm rate.
Following fundamental definition and formulas are used to estimate the performance of the classifier: accuracy rate (AR) and Error Rate (ER).
True Positive: When, the number of found instances for attacks is actually attacks.

False Positive: When, the number of found instances for attacks is normal.

True Negative: When, the number of found instances is normal data and it is actually normal.

False Negative: When, the number of found instances is detected as normal data but it is actually attack.

The accuracy of IDS classifier is measured generally on basis of following parameters:

Detection Rate: Detection rate refers to the percentage of detected Attack among all attack data, and is defined as follows:

$$\text{Detection rate} = \frac{\text{TP}}{\text{TP} + \text{TN}} * 100$$

With this formula detection rate for different types of Attacks can be calculated.

False Alarm rate: False alarm rate refers to the percentage of normal data which is wrongly recognized as attack. , and is defined as follows:

$$\text{False Alarm rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} * 100$$

VII. CONCLUSION

In this paper, Machine Learning technique have been proposed in terms of accuracy, detection rate, false alarm rate and accuracy for four categories of attack under different percentage of normal data. The purpose of this proposed method efficiently classify abnormal and normal data by using very large data set and detect intrusions even in large datasets with short training and testing times. Most importantly when using this method redundant information, complexity with abnormal behaviors are reduced. With proposed method we get high accuracy for many categories of attacks and detection rate with low false alarm. The proposed method results compare with other machine learning technique using intrusion detection to improve the performance of intrusion detection system. Experimental results and analysis shows that the proposed system gives better performance in terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

REFERENCES

[1] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[2] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[3] Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/$26.00 ©2010 IEEE

[4] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[5] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" IEEE, JANUARY 2011

[6] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[7] Jingbo Yuan , Haixiao Li, Shunli Ding , Limin Cao "Intrusion Detection Model based on Improved Support Vector Machine" Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[8] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze "A Novel Intrusion Detection Method Based on Support Vector Machines" IEEE 2010.

[9] W. Yassin, Z. Muda, M.N. Sulaiman, N.I.Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification" IEEE 2011.

[10] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques" IEEE 2010.

[11] S. SobinSoniya, S. Maria Celestin Vigila, Intrusion Detection System: Classification and Techniques 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[12] Anand Motwani, Vaibhav Patel, Anita Yadav Optimal Sampling for Class Balancing with Machine Learning Techniquefor Intrusion Detection System ISSN No. (Online): 2277-2626 (2): 47-51(2015)

[13] Mr. Sachin S. Patil, Prof. Deepak Kapgate, Prof. P.S. Prasad
A Review on Detection of Web Based Attacks using Data Mining Techniques December 2013 ISSN: 2277 128X