International Journal of Electrical, Electronics and Computer Engineering **4**(2): 47-51(2015)

ISSN No. (Online): 2277-2626

Optimal Sampling for Class Balancing with Machine Learning Technique for Intrusion Detection System

Anand Motwani*, Vaibhav Patel** and Anita Yadav***

**Assistant Professor & Head, Department Computer Science Engineering, NIRT Bhopal, (M.P.)*

***Assistant Professor, Department Computer Science Engineering, NIRT Bhopal, (M.P.)*

**M. Tech. Scholar Department Computer Science Engineering, NIRT Bhopal, (M.P.)*

(Corresponding author: Anita Yadav)

(Received 03 June, 2015 Accepted 12 August, 2015)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Information security is becoming a more important issue in modern computer system. Intrusion Detection System (IDS) as the main security defensive technique that can effectively expand the scope of defense against network 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 detection rate. Unfortunately a potential drawback of all those methods is that how to classify attack or intrusion effectively. Looking at such inadequacies, the machine learning technique on balanced classes of data is applied for obtaining the high detection rate. Also, use of internet is increasing progressively, so that large amount of data and it security is also an issue. Sampling technique is one the solution for large datasets. 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 attack classification framework based on a different model. This model also based on machine learning and sampling to improve the classification performance. The proposed work is tested on basis of Accuracy, Error rate, Detection rate and False Alarm rate. KDD CUP'99 dataset used for the approach proposed in this paper. This work suggests the framework for classification of abnormal and normal data and detect intrusions even in large datasets with short training and testing times.

Keywords*:* Sampling, Classification**,** Machine learning technique, IDS.

I. INTRODUCTION

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 [1]. New machine learning based IDS with optimal sampling is used in our detection approach. The advantage of proposed IDS (Intrusion Detection system) can greatly reduce the time for system administrators/users to analyze large data and protect the system from illicit attacks. Thus improves the performance of IDS and have low false alarm rate.

II. 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 job. 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 [1]. Let a computer program with class of tasks T is said to learn from experience E and performance measure P. If program's performance at tasks T, as measured by P, improves with experience E, we can say that program is intelligent. Thus, machine learning algorithms automatically extract knowledge from machine readable information [1].

In machine learning, computer algorithms (learners) go for automatically extract knowledge from example data sets. This knowledge can be used to generate forecasts about next or novel data generated in the future and to provide insights. Such nature of the target concepts applied to the researches at hand, this means that a computer would learn to classify alerts into incidents and non-incidents (tasks T). An expected performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly [3]. Obtained training experiences (E) could be labeled instances.

III. 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.

The authors [2] evaluated performance of a Machine Learning algorithm called Decision Tree is evaluated and compared among two other Machine Learning algorithms namely Neural Network and Support Vector Machines. The algorithms were tested on basis of accuracy, detection rate, false alarm rate for four categories of attacks. Among the experiments conducted, it was found that the Decision tree algorithm outperformed the other two algorithms. Authors compared the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

The authors [4] 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 technique used the stratified sampling and decision tree. The accuracy of proposed model is compared with presented results in order to validate the legality and accuracy of the proposed model. The results showed that the designed approach gives better and robust representation of data. The experiments and performance evaluation study of the proposed intrusion detection system are carried out with the KDD 99 dataset. The experimental results show that the proposed system achieved superior Accuracy and Low Error in identifying the records.

Kaberi Das *et al.* [5] affirmed that data sets contain very large amount of data and it is not an easy task for the user to scan the complete data set. Sampling has been often recommended as an effective tool to decrease the size of the dataset operated at some rate to accuracy. It is the process of selecting representative records which indicates the complete data set by examining a fraction.

This paper focuses on various types of sampling strategies applied on neural network. Here sampling technique has been useful on two real, integers and categorical dataset such as yeast and hepatitis data set previous to classification. Authors give the comparison of different sampling strategies for classification which gives more accuracy.

Ligang Zhou *et al.* [6] investigate the effect of sampling methods on the performance of quantitative bankruptcy prediction models on real highly imbalanced dataset. Sampling methods and quantitative models are tested on two real highly imbalanced datasets. In this work, an evaluation of model performance tested on random paired and real imbalanced sample set. The commonly used re-sampling strategies include oversampling and under-sampling. Two broadly used oversampling methods: Random Oversampling with Replication (ROWR) and Synthetic Minority Over-sampling Technique (SMOTE) are employed in this paper. Two Under-sampling sampling methods: Random under sampling (RU) and Under-Sampling Based on Clustering from Gaussian Mixture Distribution (UBOCFGMD) are also employed. It is explored that Under-sampling method is better than oversampling method because there is no major 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 closely and equally represented. Some of the sampling techniques used for balancing the datasets and the performance measures for more proper mining the imbalanced datasets are also discussed by authors. Over and under-sampling methodologies have received significant interest to answer the effect of imbalanced data sets. Sampling methods plays significant role in balancing the class distribution before learning a classifier.

The authors [9] stated 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 improved processing time and low detection rate. Consequently, Feature selection plays an imperative role while identifying the relevant features within data. Several feature selection methods are used in the work [9]. Authors compared the different feature selection methods, applied on KDDCUP'99 dataset and evaluated their performance in terms of detection rate.

IV. DATA SET AND SAMPLING

A. KDD Cup 99 dataset

In practice, we recognize that this dataset is decade old and has many criticisms for Current research. However, we believe that it is still sufficient for our experiment which aims to reflect the performance of separate machine learning approaches in a general way and find out relevant issues.

Also, the full KDD99 dataset Contain 4,898,431 records and each record contain 41 features. Due to the sufficient computing power, we do not use the complete dataset of KDD99 in the experiment but a 10% portion use of it.

This 10% KDD CUP 99 dataset contains 494,021 records (each with 41 features) and 4 categories of attacks [12]. The details of attack categories and specific types are shown in Table 1. Table 1 showing that there are four attack categories in 10% KDD99 dataset:

(1) Probe: Such attacks scan networks to gather deeper information.

(2) DoS: Denial of service denies the access to one or more resource.

(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:

Table 1: KDD Cup 99 Dataset.

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

B. Sampling

It is not easy task to for the user to scan the entire data set as it contains tremendous amount of data. The researcher's initial task is to formulate a rational justification for the use of sampling the research [5]. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy [4]. The process of selecting representative records which indicates the complete data set by examining a fraction. Sampling techniques overcomes the problems of traversing each and every element from the data base; and it is easy to study the sample rather than the entire dataset. Also, results show that it sometimes likely to produce more reliable results [3].

C. Feature selection

Due to the big amount of data flowing over the network real time intrusion detection is almost impossible. Research on Feature Selection started in early 60s [9]. It can optimally decrease the computation time and model complexity. Mainly feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features from the data for building an effective and efficient learning model [9]. A number of feature selection algorithms are proposed by different authors.

Attribute evaluator is basically used for ranking every feature according to a number of metric. Different attribute evaluators are available in WEKA. We used (Weka, 3.0.6) a learning machine tool in this job which includes CfsSubset Eval, Chi Squared Attribute Eval, Gain RatioAttribute Eval and Info Gain Attribute Eval. A CfsSubsetEval: Evaluates the value of a subset of attributes by considering the individual capability of each feature along with the degree of redundancy between them. The subsets of features that are highly correlated with the class while having low inter-correlation with the other attributes are preferred.

V. PROPOSED WORK

Researchers in machine learning community have found that the strategy of re-sampling the original dataset to deal with the issue of class imbalance is efficient. The commonly used re-sampling strategies Oversampling is to sample the minority class more and more to achieve the balanced distribution of the two classes while under-sampling is to select a portion of the majority class to accomplish the distribution balance of the two classes. In the original imbalanced training dataset, let the original set of minority class and majority class denoted by S_{min} and S_{max} separately, the size of minority class S_{min} is much less than the size of majority class S_{max}. The training dataset is denoted by S. 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 respectively. Therefore, the set of minority class S_{min} = {I1, I2} so S_{min} = 2, the set of majority class S_{max} = {M1} and so S_{max} = 1.

A. Optimal Sampling Algorithm

Let S1, S2 …Sn are classes. CE denotes class elements.

1. $\sum_{p=1}^{n} CEp = \text{Dataset Element}, \text{ where } p = 1,2,3 \ldots n.$
2. Select and make classes with Majority, Minority, Optimal & others.
3. Let S_{max} Represents Majority
   $S_{majority} = (S_{max})$, $S_{minority} = (S_{min})$
   $S_{optimal} = (S_{opt})$, $S_{others} = (S_{oth})$
4. Calculate Ratio = S_{opt}/S_{max}.
5. Under-sample the majority (S_{max}) up to optimal (S_{opt})
6. Under-sample the optimal (S_{opt}) up to other (S_{oth})
7. Apply SMOTE WITH n = 200% for minority classes SMOTE = {C_{min}, n, minority}
8. Got sampled dataset with balanced classes.

B. Under Sampling and Over Sampling by SMOTE

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 minority class becomes some specified percentage of the majority class [6-7].

In describing the experiment, our terminology will be such that if we under-sample the majority class at 200%, it would mean that the customized dataset will contain two times as many elements from the minority class as from the majority class; for example if the minority class had 50 samples and the majority class had 200 samples and we under-sample majority at 200%, the majority class would stop at having 25 samples.

Oversampling is to sample the minority class more and more to achieve the balanced distribution of the two classes [7].

By applying a mixture of under-sampling and over-sampling, the early bias of the learner towards the majority class is reversed in the favor of the minority class. Classifiers are learned on the dataset doubled by "SMOTING" the minority class and under-sampling the majority class.

Table 2: Balanced sampled KDD cup 99 dataset.

Normal	Attack				Total
	DoS	U2R	R2L	PROBE	
	97278	52	1126	4107	
4107	102563				106670

C. Architecture of the Proposed Work

In Architecture of the system shows that in 10% portion of KDD99 dataset Firstly we are applying optimal sampling technique and get balanced sampled dataset now we are using preprocessing technique in sampled 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.

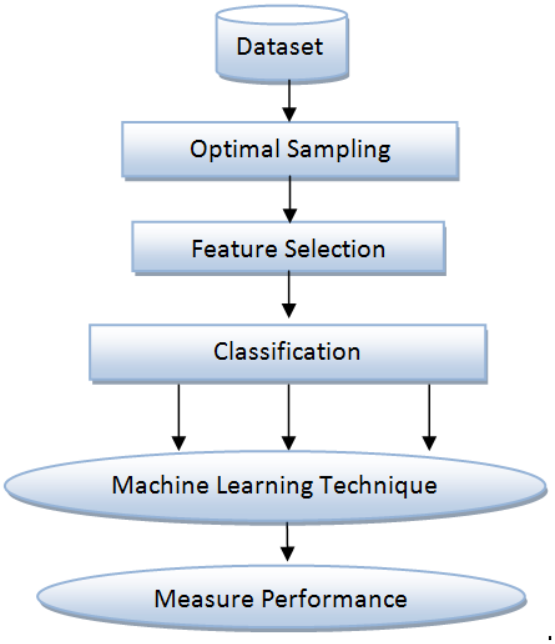

Fig. 1. Architecture of the system.

VI. RESULT ANALYSIS

Balanced sampled KDD'99 dataset, obtain from optimal sampling technique. Result shows the performance of the proposed approach classifier in terms of accuracy and error rate on sampled KDD cup 99 dataset. Result also shows comparison of 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:

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

Table 3 shows the accuracy rate and error rate of proposed method on different dataset. Proposed method has accuracy given 99.47% and 0.57 error rate, detection rate of proposed method has 81.78% and false alarm rate 0.67%. we compare the result with other classifier, and original KDD 99 dataset. Table 3 shows the result.

Table 3: Classifier Performance on Balanced Sampled Dataset.

Classifiers	Sampled dataset			
	Accuracy %	Error Rate %	Detection Rate %	False alarm Rate %
Decision Tree	99.45	0.57	81.78	0.67
Naive Bayes	99.04	0.95	79.93	1.25
Neural Network	88.80	11.13	78.91	2.56

VII. CONCLUSION

In this paper, Optimal Sampling for Class Balancing with Machine Learning Technique for Intrusion Detection System has been proposed. The proposed architecture outperforms in terms of detection rate, false alarm rate and classification accuracy for four categories of attack under different percentage of normal data. The purpose of this proposed methodology 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]. Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques", 978-1-4244-5651-2/10, 2010 IEEE.

[2]. Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 2010 *International Conference on Networking and Information Technology* 978-1-4244-7578-0, 2010 IEEE.

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

[4]. Devendra kailashiya, Dr. R.C. Jain "Improve Intrusion Detection Using Decision Tree with Sampling" Vol. **3** (3), 1209-1216 ijcta 2012.

[5]. Kaberi Das, Prem Pujari Pati, Debahuti Mishra, Lipismita Panigrahi "Empirical Comparison of Sampling Strategies for Classification" ICMOC-2012, Elsevier science direct.

[6]. Ligang Zhou," Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods." *Contents lists available at SciVerse Science Direct Knowledge-Based Systems journal homepage:* www.elsevier.com/locate/knosys online 3 January 2013.

[7]. Nitesh V. Chawla "Data mining for imbalanced datasets: an overview" sprimger.

[8]. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer "SMOTE: Synthetic Minority Over-sampling Technique" *Journal of Artificial Intelligence Research* **16** (2002) 321–357.

[9]. Megha Aggarwal, Amrita "Performance Analysis of Different Feature Selection Methods In Intrusion Detection" *International Journal of Scientific & Technology Research* Volume **2**, Issue 6, June 2013.

[10]. 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, 2010 IEEE

[11]. 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. 2010 IEEE.

[12]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. "A Detailed Analysis of the KDD CUP 99 Data Set" 2009 IEEE.