ISSN:0975-9646

Data Mining Vs Statistical Techniques for Classification of NSL-KDD Intrusion Data

Aakansha Patel, Santosh Sammarvar, Amar Naik

Department of Information technology,
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal(M.P.),India

***Abstract*- Intrusion is a kind of malicious attack and is very harmful for individual or for any organization. Due to rapid growing of internet users it has become an important research area Information and network security is becoming an important issue for any organization or individual to protect data and information in their computer network against attacks. In this study two categories of techniques :Statistical techniques and data mining technique ,one methods from each technique is considered for comparative study ,these are decision tree technique C5.0 and support vector machine (SVM) applied on widely used intrusion data i.e. NSL-KDD data set downloaded from UCI repository site. A comparative study shows that C5.0 outperformance SVM in terms of accuracy, sensitivity and specificity error measures.**

***Keywords*- Support Vector Machine (SVM), Decision Tree Technique, NSL-KDD Data.**

I. INTRODUCTION

Intrusion is a kind of malicious attack that may enter to the system or network silently. Intrusion Detection System (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces alarm to take proper and suitable action against intrusion. An IDS monitoring system must be highly efficient to protect system and network against the intrusion detection systems (IDS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. An efficient and robust Intrusion Detection System (IDS) can be developed using various techniques like data mining techniques and statistical techniques.IDS acts as a classifier which is capable to classify the data as normal or attack.

Several authors have worked and proposed many novel frameworks for classification of NSL-KDD intrusion data. Shijinn Hornj and et al. [12] have used hierarchical clustering and SVM methods to design a framework as IDS , the results are compared with other techniques and found to be satisfactory. Other others [8][9][11] have used soft computing techniques like ANN and Fuzzy logic to develop a model for NSL-KDD data [10] classification .The proposed models are performing better ,results are also reflected in terms of various statistical measures like accuracy ,sensitivity ,specificity ,ROC curve and other.

Zonghua zhang hongs hen [18] found that, modified SVMs can be trained online and the result outperform the original ones, for this they use DARPA dataset and SVM technique. Other authors [4] [5][6][13][14][18][20] have used the same KDD 99 CUP dataset with different technique and obtained d satisfactory results. Shih-wei lin and et. al proposed algorithm using SVM and decision tree techniques ,the proposed algorithm was deployed successfully in anomaly detection, one more researcher [6] has proposed a framework and found intrusion detection can prevent network intrusion, greatly improving security. Many other researchers [7][12][16][17] have used different feature selection techniques on NSL-KDD data set and investigated some novel hybrid intelligent decision technologies. levent koc and et. al [7] have used hidden naïve bayes multiclass classifier for IDS ,their results show improvement in accuracy of detecting DoS attacks.

Through deep literature survey it has been concluded that data mining and statistical techniques are widely used for development of IDS model in last on decade ,motivating from these study in this research work two different categories of techniques: Statistical and data mining based decision tree techniques have been compared to check the efficiency of IDS classifier. Models are verified with many error measures with different partitions of data. Results reveal that decision tree based technique C5.0 is performing better than statistical technique. Performance measures show that decision tree technique outperforms SVM.

II. EXPERIMENTAL DATA AND TECHNIQUES USED

One of the benchmark data widely used for intrusion detection is NSL-KDD data. This data is used to solve some of the inherent problems of the KDD'99 data set [2].There are few benchmark data available publically related to intrusion. In NSL-KDD dataset there is no duplicate records .This dataset contains number of attributes, which are supportive for measures the attacks in this we have total 5 classes and 22 sub classes. This is downloaded from repository site [10].

One of the most important applications of data mining is classification and for this decision tree based techniques are widely accepted as compare to this support vector machine (SVM) was used as machine learning tool .Many authors have applied these two techniques for classification of intrusion data. These two techniques are as follow:

- Decision tree technique as C5.0:[1]Decision tree is so popular because the construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data their representation of acquired knowledge in tree form is intuitive and

generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. Decision tree based techniques are widely used for classification of intrusion related data due to its capability of exploring knowledge in terms of simple if-then rules ,a decision tree is first inducted based on training data and then tested using testing samples.

- Support Vector Machine (SVM): Statistical models involving a latent structure often support clustering, classification, and other data mining tasks. Because of their ability to deal with minimal information and noisy labels in a systematic fashion, statistical models of this sort have recently gained popularity, and success stories can be found in a variety of applications, for example- population genetics, scientific publications, words and images, disability analysis, fraud detection, biological sequences & networks. There is a statistical model describe below-

  Support vector machines (SVM) [1] are supervised learning methods that generate input-output mapping functions from a set of labeled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyper planes are constructed to optimally separate the classes in the training data. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data by maximizing the distance between the two parallel hyper planes. An assumption is made that the larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will be.

III. EXPERIMENTAL WORK AND RESULT DISCUSSION

An experimental framework for classification of intrusion data using statistical and DT techniques is depicted in figure 1

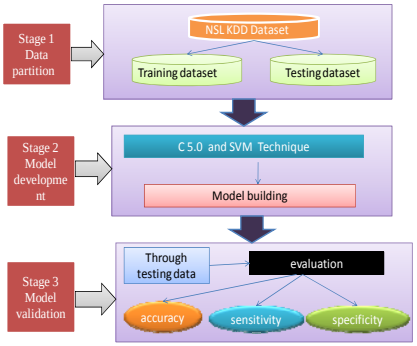

Fig. 1: An experimental framework for classification of intrusion data

There are several stages involved in this framework which are explained as follows:

(a) **Stage1 (Data Partition):** In this stage the dataset which is downloaded from UCI repository site is get partitioned into two partitions i.e. training and testing. The training samples are used for learning or training of model while testing samples are new or unseen data and are used to determine its accuracy in real sense. Partition is done randomly through the facility provided in data mining software used in this piece of research work.

(b) **Stage 2 (Model Development):** A stream is designed using Clementine data mining toll by placing various nuggets and by supplying NSL-KDD data set in CSV (Comma separated value) format. Models are trained and associated with output nugget to get the results in terms of various error measures.

(c) **Stage 3(Model Validation):** After model development, the models are evaluated or validated through following statistical measures:

Formulae and detail of measures

The formula used for calculation are-

Accuracy- TP+TN/TP+TN+FP+FN … (1)

Sensitivity- TP/TP+FN … (2)

Specificity- TN/FP+TN … (3)

Where TP= True positive

TN=True negative

FP=false positive

FN=false negative

Accuracy alone cannot verify the model in appropriate way instead we need other measures like specificity and sensitivity .All most all the authors have verified their models with these measures.

After simulation, results are obtained as shown in table 1 and 2 in case of two different partitions respectively for 50%-50% and 55%-45% as training and testing as confusion matrix, this table show number of samples classified under a particular

Category of class label .element of table contains higher number of samples which are correctly classified by the Models say for example first cell of table contains 4640 correctly classified samples under DoS category of attack while 4 and 6 samples under this category are misclassified respectively under probe and normal category. Similarly samples are classified under other category. Samples classified correctly are more in case of all the partitions and for both the techniques which show the efficiency of model as intrusion detection system with this higher true alarm rate and lower false alarm can be obtained.

A comparative results in between two techniques:SVM and C5.0 is shown in table 3 in terms of three measures calculated with the help of formula 1,2 and 3 the same is also shown in form of bar chart in figure 5.Table and figure clearly show that DT based technique is better than data mining technique.

TABLE 1. CONFUSION MATRIX IN CASE OF SVM

Partition size	Actual Vs Predicted	DoS	Probe	R2L	U2R	Normal
50-50	*DoS*	4569	0	0	0	81
	Probe	2	1053	0	0	18
	R2L	0	0	82	1	20
	U2R	0	0	1	2	2
	Normal	22	25	13	2	6731
55-45	*DoS*	4110	0	0	0	67
	Probe	1	947	0	0	18
	R2L	0	0	76	1	17
	U2R	0	0	1	2	1
	Normal	20	22	13	1	6042

TABLE 2. CONFUSION MATRIX IN CASE OF C 5.0

Partition Size	Actual Vs Predicted	DoS	Probe	R2L	U2R	Normal
50/50	*DoS*	4640	4	0	0	6
	Probe	13	1044	1	0	15
	R2L	0	2	86	0	15
	U2R	0	0	0	0	5
	Normal	5	9	10	0	6769
55/45	*DoS*	4174	1	0	0	2
	Probe	6	948	1	0	11
	R2L	0	3	79	0	12
	U2R	0	0	0	0	4
	Normal	7	8	6	0	6077

TABLE 3. THREE PERFORMANCE MEASURE FOR SVM & C 5.0

Algorithm	Partition Size	Performance Measure		
		sensitivity	Specificity	accuracy
SVM	50/50	98.25%	99.69%	98.52%
	55/45	98.39%	99.70%	98.57%
C 5.0	50/55	99.78%	99.93%	99.33%
	55/45	99.92%	99.81%	99.46%

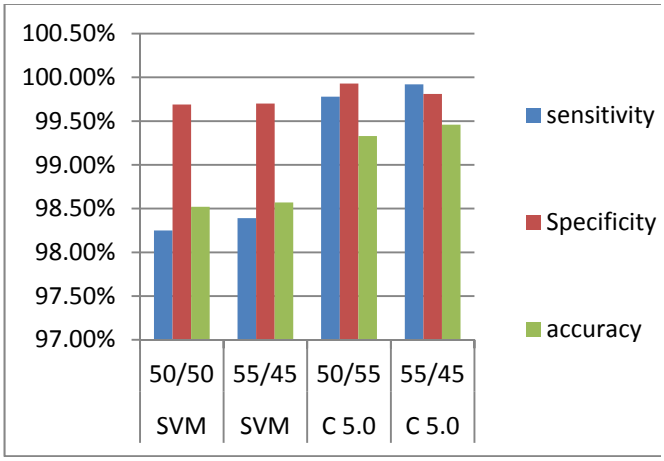

Fig 2: Performance measures of SVM & C 5.0

IV. CONCLUSION

Due to huge amount of data transmission over public network it is mandatory to protect data and information from the intruders for individual as well as for any organization. Statistical technique like support vector machine were very popular among the researchers ,many models with many variations of SVM are developed and integrated with other techniques due its internal good characteristics on the other hand relatively new technique suggested by Quinlan as C5.0 is widely accepted for development of intrusion detection system (IDS) .This study involves with a comparison of new and old techniques for intrusion related data classification based on two different partitions. An experimental result proves that C5.0 is outperforming than SVM at both training and testing stages. Accuracy, sensitivity and specificity in case of C5.0 for 55/45 partition is either better or competitive as compare to SVM.

REFERENCES

[1] Arun K. Pujari. ," Data Mining Techniques", 4th Edition, Universities Press (India) Private Limited.
[2] Gang Wang Jinxing Hao,Jian Ma,Lihua Huang,"A New Approach To Intrusion Detection Using Artificial Neural Networks And Fuzzy Clustering",Expert System With Application,2010.
[3] Krzysztof J. Cios,"Data Mining Methods For Knowledge Discovery", Kluwer Academic Publishers, 1998.
[4] Levent Koc,Thomas A. Mazzuchi Shahram Sarkani,"A Network Intrusion Detection System Based On A Hidden Naïve Bayes Multiclass Classifier",Expert System With Application,2012.
[5] Levent Koc,Thomas Shahram Sarkani,A. Mazzuchi,"A Network Intrusion Detection System Based On A Hidden Naïve Bayes Multiclass Classifer",Applied Soft Computing,2012.
[6] Lihang Yang Ni Yu,"Intrusion Detection Technology Research Based On Apriori Algorithm",Physics Procedia,2012.
[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, And Ali A. Ghorbani, "A Detailed Analysis Of The Kdd Cup 99 Data Set" Proceeding Of The 2009 Ieee Symposium On Computational Intelligence In Security And Defence Application.
[8] Mohaned M. Abd-Eldayem,"A Proposed Http Service Based Ids",Agyption Informatics Journal,2014.
[9] Mrutyunjaya Panda,Ajith Abraham,Manas Ranjan Patra,"A Hybrid Intelligent Approach For Network Intrusion Detection",Procedia Engineering,2012.
[10] Nsl Kdd Dataset Url Www.Nsl.Cs.Und.Ca/Nsl-Kdd/Kddtrain+-20persent.Txt Last Accessed On March,2014.
[11] Saurabh Mukherjee,Neelam Sharma,"Intrusion Detection Using Naïve Bayes Classifier With Feature Reduction",Procedia Technology,2012.
[12] Shi Jnn Horng,"Aa Novel Intrusion Detection System Based On Hierarchical Clustering And Support Vector Machines" 2010
[13] Shin Wei Lin,Kuo Ching Ying,Chou Yuan Lee,Zne Jung Lee,"An Intelligent Algorithm With Feature Selection And Decision Rules Applied To Anomaly Intrusion Detection",Applied Soft Computing,2012.
[14] Siva S. Sivatha Sindhu,S. Geetha,A. Kannan,"Decision Tree Based Light Weight Intrusion Detection Using A Wrapper Approach",Expert System With Applications,2012.
[15] Srilatha Chebrolu,Ajith Abraham,Johnson P.Thomas,"Feature Deduction And Ensemble Design Of Intrusion Detection System",Computers & Security,2004.
[16] SPSS Clementine help file http//www.spss.com last accessed on june 2014.
[17] V. Bolón-Canedo," Feature Selection And Classification In Multiple Class Datasets: An Application To Kdd Cup 99 Dataset", Expert Systems With Applications, 2011.
[18] Wenying Feng,Qinglei Zhang,Gongzhu Hu,Jimmy Xiangji,Hwang,"Mining Network Data For Intrusion Through

Combining Svms With Ant Colony Networks",Future Generation Computer Systems,2013

[19] Zonghua Zhang,Hongs Hen,"Application Of Online Training Svms For Real Time Intrusion Detection With Different Considerations",2005.

[20] Zubair A. Baig,Sadiq M. Sait,Abdul Rahman Shaheen,"Gmdh Based Networks For Intelligent Intrusion Detection",Engineering Application Of Artificial Intelligence,2013.