Comparison of Data Mining Techniques for Building Network Intrusion Detection Models

Harsha Kosta
Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal, India

Darshan Bhavesh Mehta
Department of Mechanical and Manufacturing Engineering,
Manipal Institute of Technology, Manipal, India

ABSTRACT
Intrusion detection is a detection of encroachment on the personal network or the private network to breach the security systems. This system provides analytical measures to gather information from various networks or computers to identify the cracks in the security systems caused by intruders. The sudden tremendous growth in the amount of internet users network intrusion detection has gained a huge amount of attention/need towards the research of network. Today, cyber-attacks have become a vital issue for any organization or individual in the network against preserving significant data and information in their personal computers connected to a network. In this paper, a comparative study was done on two different data mining techniques: decision tree and support vector machine algorithms. These techniques are implemented on the dataset for the experiment, since decision tree C5.0 technique and support vector machine (SVM) in general widely used in intrusion experiment data i.e. KDD CUP99 data set downloaded from UCI repository site. The better performance of C5.0 algorithm in terms of accuracy, sensitivity and specificity error measures are to be proved in this paper.

General Terms
C5.0 algorithm, SVM algorithm, Classification, Clustering, Network Intrusion, KDD cup99, Cyber-Attacks, DoS, Probe, R2L, U2R.

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

1. INTRODUCTION
A network intrusion is any unauthorized activity on a computer network. The instrumentation for data collection using computer network is the first step for network intrusion. monitor the network traffic is monitored and raised alarms with aid of pattern recognition which matches the traffic to a saved pattern. These alarms are categorised into serious or false positive by the security analysers. The general response to such activity is shutting this part of the network and notify the cyber-crime department for the particular intrusion for future reference. The data analysts then analyse the pattern recorded for the intrusion using their signatures, these results might not be meaningful if the network is too large but on small networks it works the best. The reason why the large network is a problem is that in a network there are almost over a million alarms raised in a day which keeps increasing as the network expands its boundaries due to which the pattern recognition cannot select the most significant pattern among the various other patterns detected. Commercial tools usually do not grant any enterprise level view of alarms raised by vendor's sensors. These software packages for intrusion detection are mostly signature-oriented with almost or no state information on record. These limitations guided us to investigate and use the techniques of data mining to solve this problem. Many authors have studied and proposed many unique models for classification of NSL-KDD intrusion data. Shijinn Hornj and et al. [12] have used hierarchical clustering and SVM methods to design model in NIDS, the results are compared with other techniques and found to be satisfactory. Others [8], [9] and [11] have used soft computing techniques like ANN and Fuzzy logic to develop a model for NSL-KDD data [10] classification. Levant koc and et. al [7] have used hidden Naïve Bayes multiclass classifier for NIDS, their results are an improvement in accuracy of detecting DoS attacks. Models are verified with many error measures and using various partitions of data. Results reveal that decision tree based technique C5.0's performance is comparatively better than SVM. Based on performance, decision tree technique outruns SVM.

2. EXPERIMENTAL PROCEDURE
2.1. Description of Experimental Data and Techniques Used
A standard dataset is used for network intrusion. This dataset has been used since it was represented in The Fifth International Conference on Knowledge Discovery and Data Mining. This data is used to solve some of the inherent problems of the KDD'99 data set [2]. In NSL-KDD dataset there is no duplicate records. A number of attributes in this dataset making measuring the attacks in easy by having a total of five classes and 22 sub-classes. This dataset was obtained from the website [10].

The most important data mining technique is classification which is used for the comparison by many authors for network intrusion data are machine learning tool (SVM) and the decision tree algorithms.

The decision tree algorithm C5.0 is one of the most widely used and practical methods for inductive inference over supervised data. It is a method for classifying categorical data depending on their attributes. The efficiency of this algorithm is also shown while analysing huge amount of data, so it is widely used for data mining applications. The feature that makes this appropriate for exploratory knowledge discovery is that its construction does not need the domain knowledge or parameter

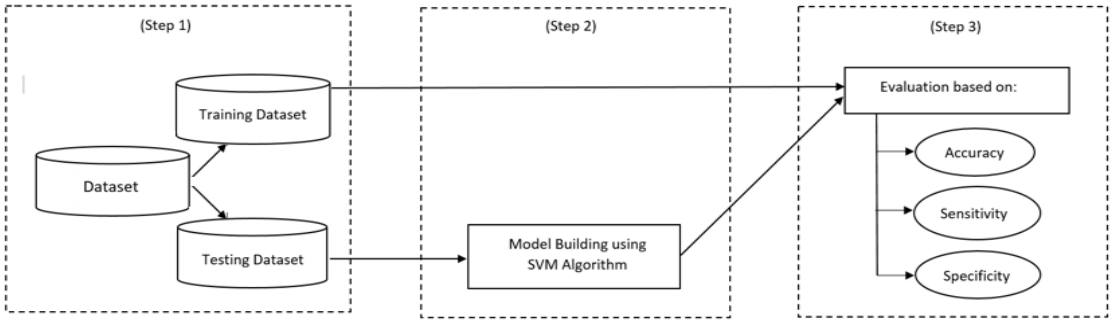


Figure 1: An operational framework depicting classification of intrusion data.

setting. The representation of acquired knowledge in this form is more inductive and comprehendible by humans. The characteristic of being fast, simple, accurate make the learning

and classification of decision tree a classic technique used for intrusion data. Also the ability of exploring knowledge in terms of simple if and else rules this algorithm will be conscripted on training data first and on the testing dataset.

SVM (Support Vector Machine) is a classification technique is a supervised learning technique which analyses data and recognizes and generates patterns from them. From statistical learning theory, SVM a machine learning algorithm is derived. SVM classification use a very small sample set and generate pattern from that. For classification, the commonly used methods are nonlinear kernel functions that transform the input data to a high dimensional feature space where the input data can be finally separated.

2.2. Methodology and Result Discussion

An operational framework of classification using data mining techniques has been depicted in figure 1. The following are the steps involved in building the classification model for intrusion:

Step 1 (Partition): The original dataset is obtained from the UCI repository site. The dataset is partitioned into two sets training data set and testing dataset. The partition is randomised through using a data mining software particularly used for this sort of research work. After the partition the training dataset is used to learn and train the model while the

testing data is new and unseen data used to test the accuracy of the model.

Step 2 (Model building): Clementine data mining tool is used to design a stream by supplying NSL-KDD data set in CSV (Comma separated value) format and placing nuggets in the dataset. Models are trained and to output a nugget to get the various error measures as results.

Step 3 (Testing of the model): Once the module was successfully built the model was assessed on certain statistical criteria having the formula –

$$\text{Accuracy: } P_t + N_t / P_t + N_t + P_f + N_f + \ldots(a)$$
$$\text{Sensitivity: } P_t / P_t + N_f \ldots(b)$$
$$\text{Specificity: } N_t / P_f + N_t \ldots (c)$$

Here, P_t = True positive
N_t = True negative
P_f = false positive
N_f = false negative

Since taking account of the only factor "accuracy" is not enough also the factors like specificity and sensitivity were also taken into account, these factors have been used by all the other authors as well in case of NID.

After the accomplishment of the experiment, the results which are obtained are shown in the matrices represented in the form of a tables 1, 2, 3 and 4.

Table 1: For partition size 50% resulting matrix for SVM technique

Actual vs predicted	DoS	Probe	R2L	U2R	Normal
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

Table 2: For partition size 55%-45% resulting matrix for SVM technique

Actual vs predicted	DoS	Probe	R2L	U2R	Normal
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

Table 3: For partition size 50% resulting matrix for C5.0 technique

Actual vs predicted	DoS	Probe	R2L	U2R	Normal
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	2	6042

Table 4: For partition size 55%-45% resulting matrix for C5.0 technique

Actual vs predicted	DoS	Probe	R2L	U2R	Normal
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

These results are taken into account using the two partitions respectively for 50%-50% and 55%-45% as training and testing as a matrix, these tables show the amount of samples classified under each category. Each element of table depicts that higher

number of samples which are correctly classified by the model.

	SVM	C5.0	
Sensitivity	98.25%	99.78%	Partition size 50%
accuracy	99.69%	99.93%	
specificity	98.52%	99.33%	
Sensitivity	98.39%	99.92%	Partition size 55-45%
accuracy	99.70%	99.81%	
specificity	98.57%	99.46%	

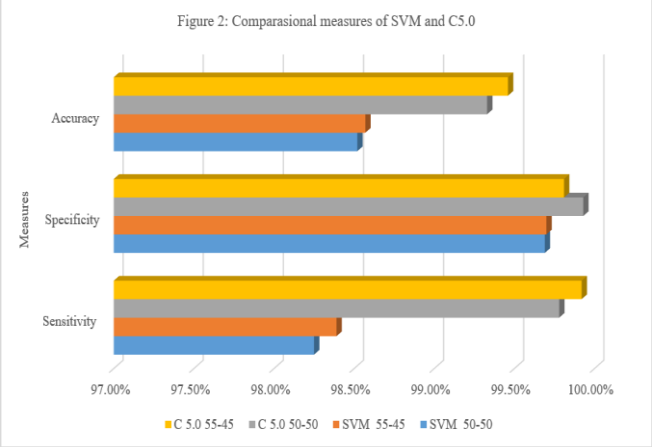
Figure 3: Comparison table for C5.0 and SVM

For example, first cell of table 2 contains 4640, which means that 4640 samples have been correctly classified under DoS category of attack while 67 samples under have been misclassified respectively under the normal category. Similarly, the samples have been tabled in the rest. The more the correct classification of these samples the more the efficiency of these algorithms are. This also shows that the model can obtain higher genuine alarms and lower false alarms in intrusion detection. A comparative result of the two techniques, SVM and C5.0 with respect to the three measures calculated with the help of formula a, b and c is depicted by Table 5 and the bar graph, Figure 2. This clearly states that C5.0 is a better technique for classification of the intrusion detection data.

3. CONCLUSION

The huge amount of data transaction in today's world over the common public network calls for protection of data and information from intruders for every type of user, be it an organisation or an individual. Techniques like SVM were widely used by the authors for building data models to prevent intrusion or detect intrusion due to its better characteristics over the rest of the algorithms. Contradicting to this, a new technique proposed by Quinlan recently got accepted over a huge population for the development of intrusion detection system.

Since both the techniques were so popular among the researchers, this study is to compare the new and old methods to build a better network intrusion system. An experimental result, proves that C5.0 technique outruns the SVM techniques based on accuracy, specificity and sensitivity with the partition of 50%- 50% and 55% -45%.



4. ACKNOWLEDGEMENT

Manipal Institute of Technology, Manipal has not contributed for or towards the project in any way and in no way has supported us in pursuing this research.

5. 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] Lihang Yang Ni Yu," Intrusion Detection Technology Research Based On Apriori Algorithm", Physics Procedia,2012.

[6] 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.

[7] Mohaned M. Abd-Eldayem,"A Proposed Http Service Based Ids", Agyption Informatics Journal,2014.

[8] Mrutyunjaya Panda,Ajith Abraham,Manas Ranjan Patra,"A Hybrid Intelligent Approach For Network Intrusion Detection", Procedia Engineering,2012.

[9] Nsl Kdd Dataset Url Www.Nsl.Cs.Und.Ca/Nsl-Kdd/Kddtrain+-20persent. Txt Last Accessed On March,2014.

[10] Saurabh Mukherjee,Neelam Sharma, "Intrusion Detection Using Naïve Bayes Classifier With Feature Reduction",Procedia Technology,2012.

[11] Shi Jnn Horng,"Aa Novel Intrusion Detection System Based On Hierarchical Clustering And Support Vector Machines" 2010

[12] 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.

[13] Siva S. Sivatha Sindhu,S. Geetha,A. Kannan,"Decision Tree Based Light Weight Intrusion Detection Using A Wrapper Approach",Expert System With Applications,2012.

[14] Srilatha Chebrolu,Ajith Abraham,Johnson P.Thomas,"Feature Deduction And Ensemble Design Of Intrusion Detection System", Computers & Security,2004.

[15] SPSS Clementine help file http//www.spss.com last accessed on June 2014.

[16] V. Bolón-Canedo," Feature Selection and Classification in Multiple Class Datasets: An Application to Kdd Cup 99 Dataset", Expert Systems with Applications, 2011.

[17] 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

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

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