Comparative Study of Different Models before Feature Selection and AFTER Feature Selection for Intrusion Detection

Janmejay Pant
Graphic Era University

Bhaskar Pant, Ph.D
Graphic Era University

Amit Juyal
Graphic Era University

ABSTRACT
A network data set may contain a huge amount of data and processing this huge amount of data is one of the most challenges task for network based intrusion detection system (IDS). Normally these data contain lots of redundant and irrelevant features. Feature selection approaches are used to extract the relevant features from the original data to improve the efficiency or accuracy of IDS. In this paper an effective feature selection approaches are used for the NSL KDD data set. The performance of the used classifiers measure and compared with each other.

KEYWORDS
Feature selection, intrusion detection, NSL-KDD, Weka

1. INTRODUCTION
An intrusion can be defined as a series of action aiming at compromising the security of a computer network system [1]. Intrusion may affect the integrity, availability of the computer recourse. Intrusion may be external attacks or internal attacks [2]. The process of detecting and preventing intrusion activity is known as intrusion detection. It a significant way of defending the computer system from intrusions. There are two kinds of intrusion detection one is host based and another one is network based. Host based IDSs examine the internal data of computer system and network based IDSs examine data exchanged between computers [3].

Classification and clustering are two techniques which have been applied in intrusion detection system. Classification is learning a function for categorizing unseen data into one of several predefined classes based on training set. Clustering is a technique where the classes are not predefined at the stage of learning. Both methods can be used for intrusion detection. If our purpose is to distinguish the abnormal from the normal action then classification is more appropriate to accomplish the task. If the purpose of system seeks to identify the type of attack clustering is suit [4].

Feature selection is used to reduce time and increase the accuracy of the system i.e. it gives the high detection rate and low false alarm rate.

Feature selection is a process of selecting relevant features and removing irrelevant or redundant features from the original data set which plays an important role in many different areas such as statistical pattern recognition, machine learning, data mining and statistics [5].

In this paper, an effective feature selection method is applied in network intrusion detection. By a detailed comparison with other used methods. Finally we compare the performance of all the approaches with respect to time taken by the classifier and in terms of accuracy of the classifier. Our proposed can successfully recognize the important feature selection to building IDS.

2 .RELATED WORK
In many areas of machine learning feature selection has been applied. For some concepts, all features are important, but for some target concepts, only a small subset of features is normally relevant [6]. It was interesting to conclude that as the number of applied features was less than a specific number [7] and it gives significantly better performance. As far as intrusion detection is concerns, features selection achieves two main goals, the first one is it helps to decrease computing time by reducing the dimension of data collected by Intrusion Detection System. The second goal is, it selects high quality features which makes IDSs retain high detection rate and low false alarm rates.

Models of features selection are generally into two categories named as filter and wrapper approach [8]. In filter approach each feature evaluates first independently from the classifier then ranks the features after evaluation and keeps the superior one [9]. This may be done by using distance measure, dependency and statistics [10]. But in wrapper approach each feature or feature subset is evaluated by a classification algorithm [11]. Wrapper methods contains inbuilt algorithm for feature selection and the subset for which the classification algorithm has the best performance is selected. In general, the speed of wrapper approach is slower than the filter approach because there are repeated iterations and cross validation to evaluate the subset. But it seems that wrapper approach is more reliable because classification algorithm affects the accuracy, although the selection of the subsets an NP- hard problem.

Currently both the above techniques (filter and wrapper) are applied in intrusion detection. If we are talking about filter approach Principal Component Analysis (PCA) is used to reduce the high dimensional data by selecting features corresponding to the highest Eigen values [12]. You Chen et al. [13] introduced two filter approaches, named CFS and PCA, and two wrapper approaches, named SVM and Genetic Algorithm. Yang Li et al. [14] proposed an intrusion detection model that was efficient for feature selection. Gary Stein et al. [15] used a GA based feature selecting algorithm. This algorithm is based on wrapper approach of feature selection. The evaluation component was a decision tree and search component was GA in this algorithm proposed by Gary Li et al. Features may be important, very important and may be unimportant based on the criteria. Andrew H.Sung et al. [16] used Support Vector machines and neural networks to classify the importantance of features considering the three main criteria- the very first is over all accuracy of classification, False Positive Rate and False Negative Rate.

3. EXPERIMENTS AND RESULTS

The different feature selection algorithms are used and analyze the results. Our experiment is based on NSL-KDD data set of intrusion detection. In this part we used four methods IG [17], GR [18], Relief [19] and ChiSquare [20] for methods for feature selection. And then compare the performance of all the methods.

NSL-KDD data set is used to perform the experiments through the WEKA. It consists of a good and reasonable proportion of various types of records [21]. Each record in dataset contains forty-one attributes and one class or decision attribute. The different types of features are-

Duration,Protocol_type,Service,flag,src_bytes,dst_bytes,land, wrong_fragment,urgent, hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root, num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_host_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host__srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate, classes.

3.1Experiments Result

Firstly we use the approaches to select the important features which can help to improve the accuracy of classification. Then we get the features selected by each technique.

Feature Selection Technique	Noof Selected features	Selected features
GainRatioAttributeEval	35	12,26,25,4,6,5,30,29,3,34,33,8,35,23,31,32,16,28,27,15,2,10,13,19,1,18,17,24,14,22,21,11,9,7,20
Principal Components Analysis	83	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83
InfoGainAttributeEval	35	5,3,6,4,30,29,33,34,35,12,25,23,26,32,31,24,2,27,28,1,10,8,13,16,19,22,17,15,14,18,21,11,7,9,20
ChiSquaredAttributeEval	35	5,3,6,4,30,29,33,34,35,12,23,25,26,32,31,24,2,27,28,1,10,8,13,16,19,22,17,15,14,18,21,11,7,9,20

The feature selection is used in WEKA tool and we find the different number of features selected by different types of feature selection techniques. After feature selection the data is classified by different classifiers and then we compare the performance of the classifiers based on various parameters like accuracy, time taken by the classifier etc. we are going to discuss the performance of the classifiers before feature selection and after feature selection. In our data set there are 23866 instances and 36 attributes. We select 10 cross validation test mode for our experiment. The selected features are shown in table 1.

Table 2 Before Feature Selection

Technique	Time Taken	Correctly Classified Instances		Incorrectly Classified Instances		Accuracy TP Rate	Accuracy FP Rate
ZeroR	0.03 Sec	12738	53.373%	11128	46.627%	100%	100%
OneR	0.57 Sec	22973	96.2583%	893	3.74175	94%	1.2%
BayesNet	1.48 Sec	22720	95.1982%	1146	4.8018%	98.8%	8.9%
NaiveBayes	0.33 Sec	21522	90.1785%	2344	9.8215%	93.2%	13.3%
J48	3.17 Sec	23768	99.5894%	98	0.4106	99.7%	0.6%

Table 3 After Feature Selection

Technique	Time Taken	Correctly Classified Instances		Incorrectly Classified Instances		Accuracy TP Rate	Accuracy FP Rate
ZeroR	0.01 Sec	12738	53.373%	11128	46.627%	100%	100%
OneR	0.05 Sec	22973	96.2583%	893	3.74175	94%	1.2%
BayesNet	0.14 Sec	22908	95.9859%	958	4.0141%	94.8%	2.7%
NaiveBayes	0.04 Sec	20780	87.0695%	3086	12.930%	95.8%	2.2.%
J48	0.43 Sec	23669	99.1746%	197	0.8254%	99.6%	0.3%

We use the commonly used method to perform feature selection on NSL-KDD data set. Each of five methods is combined with the ranker search method. To evaluate the methods there are many measures for calculating the performance such as True Positive rate (TP) and false Positive Rate (FP) [22].

3.2Analysis of Result

Before the feature selection i.e. when full data set is used the accuracy of each method is mentioned above table 2. In this process zeroR is takes minimum time to build. The correctly classified instance by the approach of J48 and the accuracy of J48 is highest. But after removing some features from the data set the accuracy of each method is mentioned in Table 3. And it shows that after feature selection the accuracy of the method is increased. And time taken to build the model is also decreased. The j48 method can achieve 99.6% classification accuracy which is the highest one among all the methods.

4 CONCLUSIONS

In this paper, we have applied the various classification approaches for intrusion detection. In order to evaluate the performance of each method we used full data set first time then find the accuracy.And then a feature selection is applied and evaluates the performance of each method and again find the accuracy. A detailed comparison among the all methods is conducted on the NSL-KDD data set. Experimental results illustrate that the J48 method has the highest accuracy than other method

5 REFERENCES

[1] R. Agarwal and M. V. Joshiy, PNrule: A new framework for learning classifier models in data mining (a case-study in network intrusion detection) ,Citeseer2000.

[2] M. Sheikhan, et al., Application of Fuzzy Association Rules-Based Feature Selection and Fuzzy ARTMAP to Intrusion Detection, Majlesi Journal of Electrical Engineering, vol. 5, 2011.

[3] R. Chattemvelli and R. Sridevi, GA Approach for Network Intrusion Detection, International Journal of Research and Reviews in Information Sciences (IJRRIS), vol. 1, 2012.

[4] M. Kantardzic, Data mining: concepts, models, methods, and algorithms: Wiley-IEEE Press, 2011.

[5] L. Yu and H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy,Journal of Machine Learning Research, vol. 5, Dec. 2004, pp.1205 -1224.

[6] P. Patil, V. Attar, Intelligent Detection of Major Network Attacks Using Feature Selection Methods,Proc. International Conference on Soft Computing for Problem Solving (SocProS 2011),Springer Press, Dec. 2011, pp. 671-679, doi: 10.1007/978-81-322-0491-6_61.

[7] M. Y. Su, K.C.Chang, H.F. Wei, and C.Y. Lin, Feature Weighting and Selection for a Real-Time Network Intrusion Detection System Based on GA with KNN, Intelligence and Security" June 2008

[8] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Boston:Kluwer Academic, 1998.

[9] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, vol. 3, Jan.2003, pp.1157 -1182.

[10] Y. Chen, Y. Li, X.Q. Cheng, and L. Guo, Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System, Information Security and Cryptology, vol. 4318,Dec.2006,pp.153-167.

[11] T.M. Chen, X.M. Pan, Y.G. Xuan, J.X. Ma, and J. Jiang, A Naïve Feature Selection Method and Its Application in Network Intrusion Detection, Proc. International Conference on Computational Intelligence and Security(CIS '10), IEEE Press, Dec 2010, pp.416-420,doi: 10.1109/CIS.2010.96.

[12] G.R.Zargar and P.Kabiri, Selection of Effective Network Parameters in Attacks for Intrusion Detection, Advances in Data Mining. Applications and Theoretical Aspects: Lecture Notesin Computer Science, vol.6171,July 2010.

[13] Y. Chen, Y. Li, X.Q. Cheng, and L. Guo, Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System, Information Security and Cryptology, vol. 4318,Dec.2006,pp.153-167.

[14] Y. Li, B. Fang, Y. Chen, and L. Guo, A Lightweight Intrusion Detection Model Based on Feature Selection and Maximum Entropy Model, Proc. International Conference on Communication Technology(ICCT '06), IEEE Press,Nov.2006.

[15] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection, Proc. the 43rd annual Southeast regional conference, ACM, Mar2005, pp.136-141, doi: 10.1145/1167253.1167288.

[16] A. H.Sung and S. Mukkamala ,Identifying Important Features for Intrusion Detection using Support Vector Machines and Neural Networks, Proc. Symposium on Applications and the Interne (SAINT'03),IEEEPress,Jan.2003.

[17] L. Yu and H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research, vol. 5, Dec. 2004, pp.1205 - 1224.

[18] J.R. Quinlan, C4 .5: Programs for Machine Learning, Morgan Kaufman, 1993.

[19] K. Kira and L. A. Rendell, A Practical Approach to Feature Selection, Proc. Ninth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., 1992, pp. 249-256. 10.1007/3 540-57868-4_57.

[20] H. Liu and R. Setiono, Chi2: Feature Selection and Discretization of Numeric Attributes, Proc.Seventh International Conference on Tools with Artificial Intelligence(TAI '95) , IEEE Press,Nov.1995,pp.388-391.

[21]Tavallaee, M.Bagheri,E.,Lu,Wei.,Ghorbani,A.A.: A detailed Analysis of the KDD CUP 99 Data Set ,In: the Proceedings of the 2009 Symposium on Computational Intelligence in Security and Defense Application,2009.

[22] T.Fawcett, An Introduction to ROC Analysis, PatternRecognitionletters,Vol.27,June 2006,pp.861-874.