Combinational Feature Selection Approach for Network Intrusion Detection System

Tanya Garg
Centre for Computer Science & Technology
Central University of Punjab, Bathinda
India
Tanyagarg5023@gmail.com

Yogesh Kumar
Department of Computer Science and Engineering
Punjab Institute of Technolgy, Kapurthala
India
Yksingla37@gmail.com

Abstract-**In the era of digital world, the computer networks are receiving multidimensional advancements. Due to these advancements more and more services are available for malicious exploitation. New vulnerabilities are found from common programs and even on vulnerability in a single computer might compromise the network of an entire company. There are two parallel ways to address this threat. The first way is to ensure that a computer doesn't have any known security vulnerabilities, before allowing it to the network it has access rights. The other way, is to use an Intrusion Detection System. IDSs concentrate on detecting malicious network traffic, such as packets that would exploit known security vulnerability. Generally the intrusions are detected by analyzing 41 attributes from the intrusion detection dataset. In this work we tried to reduce the number of attributes by using various ranking based feature selection techniques and evaluation has been done using ten classification algorithms that I have evaluated most important. So that the intrusions can be detected accurately in short period of time. Then the combinations of the six reduced feature sets have been made using Boolean AND operator. Then their performance has been analyzed using 10 classification algorithms. Finally the top ten combinations of feature selection have been evaluated among 1585 unique combinations. Combination of Symmetric and Gain Ratio while considering top 15 attributes has highest performance.**

***Index terms*—Intrusion Detection System, NSL-KDD Dataset, WEKA, Data Mining, Feature Selection Techniques, Garret's Ranking Technique, Boolean AND operator**

I. INTRODUCTION

According to (Elngar et al., 2013), a computer network intrusion is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a network resource. In general, intrusion attempts are malicious actions that have the purpose of intentionally violating the system security properties. An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activities and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network.To classify the anomalous network traffic into intrusions or normal, intrusion detection model is required and that intrusion detection model analyses network traffic on the basis of 41 attributes which is a huge number of attributes. Some of these attributes are irrelevant and gives no information. There is need to reduce the number of such attributes so that intrusions can be accurately detected in short period of time. This work aims to find the optimal set of evaluation attributes so that the performance of intrusion detection model can be improved.

Selection is an important data pre-processing tool in the field of Data Mining. The research is in the growing stage in this field for the past three decades [2]. As the data is growing day by day on network in terms of features and instances, it is necessary to reduce that data to reduce its processing time and to achieve higher accuracy in results. This process of removing irrelevant data is called Data Mining. In our research work we have to reduce the features for the Network Intrusion Detection Dataset, for which we used the ranking based feature selection techniques along with ranker to get the features ranked according to their importance. To analyze intrusions 41 attributes are considered. Our aim is to reduce number of those attributes so that intrusions can be detected in short period of time with higher accuracy. The use of feature selection techniques is to remove the irrelevant or useless features that are not contributing any useful information but are wasting time of intrusion detection model. There are several performance metrics to evaluate performance of Feature Selection Techniques. In this work, NSL-KDD (Network Security Layer-Knowledge Discovery in Database) compatible classification algorithms have been evaluated using WEKA (Waikato Environment for Knowledge Analysis) tool. Then Top ten [26] classification algorithms are used to evaluate the Six Ranking Based Feature Selection Techniques. The performance of the Feature selection Techniques have been measured by considering all the performance metrics: Accuracy, ROC (Receiver Operational Characteristics), Training time, FPR (False Positive Rate), Precision, Recall, Mean absolute error and Kappa.

In this paper, initially in section II, WEKA tool has been discussed, then ten classification algorithms to be used have been discussed in section III. Various Feature selection techniques have been discussed in section IV. Chosen

dataset has been introduced in section V. Performance metrics to be used are discussed in section VI.Results and discussions have been done in Section VII and conclusions are discussed in Section VIII.

II. WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS (WEKA)

WEKA is a Data Mining, Machine Learning Tool [1]. It is a collection of number of Machine Learning and Data Mining algorithms. This software is written in Java language and contains a GUI Interface to interact with data Files. It contains 49 data pre-processing tools, 76 classification algorithms, 15 attribute evaluators and ten search algorithms for feature selection [2]. It contains three algorithms to find association rules. It has three Graphical User Interfaces: "The Explorer", "The Experimenter" and "The Knowledge Flow." The WEKA supports data stored in ARFF (Attribute Relation File Format) file format. WEKA provides the capability to develop and include the new Machine Learning algorithm in it. The algorithms can be directly applied to dataset [7].

III. CLASSIFICATION ALGORITHMS

Classifiers are also called as machine learning algorithms or classification algorithms. They are used to classify the network traffic when used on dataset. They are responsible for classification of network traffic as normal or intrusive, if intrusive then to which category it belongs.

In our work there are 72 classifiers that are compatible on our chosen dataset and on the basis of their performance we have evaluated best ten classifiers and then these classifiers have been used for feature selection process. In this section the ten classifiers we are using for our work are discussed as:

Lazy classifiers: These learners are called as lazy learners also because they are computationally expensive i.e. they take a lot of time for computation for classification as their most of the power resides in the matching scheme. It takes large amount for classification due to the reason that each example to be classified must be compared to the each of the example in training dataset. The lazy classifiers we used are lazy IB1 and IBk [4].

Random Forest: It is an ensembling classification algorithm and is based upon decision tree algorithm and produces output in the form of individual trees. This algorithm is a combination of bagging idea and random selection of features to construct a collection of decision trees with controlled variation. It is one of the highest accurate classifier for many datasets. A large number of variables can be handled by this algorithm without ignoring any variable [22].

Random Tree: Random Tree as its name indicating it's a tree build by picking random branches from a possible set of trees. Each tree has an equal probability of being get sampled in this algorithm or we can say the trees are distributed in a uniform way. Random Trees can be generated easily and efficiently. Combination of large sets of Random trees mostly designs accurate models [17].

JRip (Extended Repeated Incremental Pruning) is that type of rule based classifier that implements a propositional rule that performs repeated pruning to reduce errors and also called as RIPPER (Repeated Incremental Pruning to produce Error Reduction). JRip is a rule learner that exactly works like commercial rule learner RIPPER [4].

NB Tree: NB Tree (Naïve Bayes) is used and applicable to scale large databases and used to improve the performance of decision trees and Naïve Bayesian Classifiers. Attribute need not to be independent for this classification algorithm [12].

Rotation Forest: Rotation Forest is also an ensemble classifier that transforms the data, subset of instances, subsets of classes and subsets of features using Principal Component Analysis because this method of transforming data requires less storage space and has low computation time. Rotation Forest is based upon two base classifiers: decision tree and Forest. It works on two key components: diversity and accuracy [5].

IV. FEATURE SELECTION

Feature selection is the process of choosing relevant features from large number of features for a particular dataset by applying particular evaluation criteria as desired by user. Generally Feature selection process involves three phases. It starts with selecting subset of original features and then it evaluates the worth of each feature for that particular dataset as we are using NSL-KDD Dataset for Intrusion Detection. Then after evaluation of worth of each feature by learning algorithms, the features having lower rank or lower value can be eliminated from dataset. Finally the third step is to Evaluate the reduced feature set with classification algorithms to check whether it gives better result than previous one or not. If yes then that will be the final efficient reduced dataset [11].

Feature selection is broadly classified in two categories: one is Feature subset selection and other is feature ranking. The techniques used for Feature subset selection are different from that of the feature ranking techniques. Feature subset selection techniques gives the best reduced feature set according to algorithms used for searching best feature subset [21]. They do not rank the features. Feature ranking techniques calculates the score of each feature and rank them accordingly and we can pick the top k (as required by user) features according to higher rank and ignoring those having lower rank. We here are using Feature ranking for feature selection in our work Now the Feature selection can be done using three models i.e. one of these models can be used for feature selection. Feature selection can be done using Filter Model, Wrapper Model or Hybrid Model. Filter model performs feature selection technique without use of learning algorithm while Wrapper model involves use of learning algorithm first and then performs feature selection. Hybrid model involves use of both of these models together for feature selection.

Our work is focused and limited to filter model using seven ranking based feature selection techniques. The main advantage of using filter model is that it work without using learning algorithm, so it is unbiased. Secondly it is very simple and easy to use. It is very easy to design algorithms using his model due to its simple structure.

Information Gain (IG): Information Gain feature selection technique is based on the concept of entropy. Entropy is a measure of how pure or impure a variable is (Moore. W.A.). The expected resultant value for the feature by using information gain method is the mutual information of target variable say X and independent variable say Y. It is the reduction in the entropy of a target variable (X) achieved by using Learning algorithm (Y). Information gain method has a drawback that it chooses attributes having large distinct values over the attributes having fewer distinct values even though they are more informative [19].

Gain Ratio (GR):The Information Gain is biased towards tests with many outcomes [22]. Gain Ratio is a modification in Information gain to reduce its biasness. It takes into account the number of branches while choosing an appropriate attribute. It is based on the information given by intrinsic attributes. In this method the value of attribute decreases as intrinsic information gets larger. Gain ratio chooses an attribute only when its intrinsic information is very low. Intrinsic information means the amount of information which is needed to decide which branch belongs to which instance (Frank & Witten, 2011). The attribute having maximum value of gain ratio will be selected as the splitting attribute.

ReliefF:The Feature Selection Technique ReliefF was proposed by Kira and Rendell in 1994. It is very easy to use and is fast and accurate technique [23]. Relief works by measuring the ability of an attribute in separating similar instances. It can also be used to deal with noisy data and regression problems.

The process of selecting and ranking the features by using this technique involves basically three steps:

1. Calculate the nearest miss and nearest hit.
2. Calculate the weight of a feature.
3. Return a ranked list of features or the top k features according to a given threshold.

One R: It is a rule based algorithm as it generates rules and on the basis of those rules it selects features and rank them accordingly. OneR constructs rules and tests a single attribute at a time and branch for every value of that attribute.

Symmetrical Uncertainty:It is a correlation based Feature Selection approach. Correlation based feature selection evaluates the merit of a feature in a subset using a hypothesis – "Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other" [22]. This Feature selection method is used to measure the degree of association between the discrete attributes. It is also based on the concept of entropy and is a symmetric measure to measure correlation between set of features.

Filtered Attribute Evaluator: It is based on the same principle of Information Gain Feature Selection Technique. In our work we used this technique and found it gives the attributes in the same order as selected by Information Gain Method.

Chi-Square: It is based on the statistical theory. It measures the lack of independence between the attributes [22]. It can test strength of relationship between two variables. It predicts the value of an attribute using observed and expected values of attributes. The value of an attribute based upon this feature selection method can be calculated using the formula:

$$\chi^2 = \sum_{ij} (O_{ij} - E_{ij})^2 / E_{ij} \qquad (1)$$

Here i and j are two attributes and O means observed value and E means expected value and means value of chi-square. Higher the value of chi-square of an attribute higher will be its rank.

V. DATA SET DESCRIPTION

The dataset used for experimental evaluation consists of randomly selected instances from NSL-KDD Dataset [5]. It classifies network traffic in five categories. The number of each class instances included in training and testing dataset are mentioned in table-I.

TABLE I. DESCRIPTION OF DATASET

Class Type	Instances in Training Dataset	Instances in Testing Dataset
Normal	48522	1478
Dos	41699	37490
Probe	3608	8301
U2R	21	28
R2L	50	1075

VI. PERFORMANCE METRICS

The performance metrics that have been used to evaluate the performance are discussed as below:

Accuracy: It is the percentage of correct predictions. On the basis of Confusion Matrix it is calculated by using the formula below:

Accuracy= TP+TN/n Here n is total number of instances.

Mean Absolute Error: It is the mean of overall error made by classification algorithm. Least the error and best will be the classifier.

TPR: True Positive Rate is same as accuracy so we have not considered this metrics.

FPR: False Positive Rate is calculated by using the formula:

FPR=FP/TN+FP

Recall: It is the proportion of instances belonging to the positive class that are correctly predicted as positive.

Recall=TP/TP+FN

Precision: It is a measure which estimates the probability that a positive prediction is correct

Precision=TP/TP+FP

Training time: It is the time taken by Classifier to build the model on dataset. It is usually measured in seconds.
Kappa: Its value ranges from 0 to 1. 0 means totally disagreement and 1 means full agreement. It checks the reliability of Classifying algorithm on dataset.
ROC (Receiver Operating Characteristics): It is used to design the curve between TPR and FPR and the area under curve is called as AUC gives the value of ROC. More the area under curve and more will be the value of ROC.
Equal weightage has been given to all the performance metrics and performance of various combinations of feature selection techniques has been evaluated on the basis of scores assigned to them by Garret's Ranking Technique

VII. EXPERIMENTAL SETUP AND RESULTS

All the experiments have been evaluated on selected instances from NSL-KDD Dataset using ten classification algorithms using 5 cross validations.These classifiers have been evaluated using WEKA Data mining tool Version 3.6.10. Experiments have been performed to compare the performance of combinations of six ranking based feature selection techniques: Info gain, Gain Ratio, ReliefF, Chi-square, Filtered attribute evaluator and Symmetrical Uncertain (Filter Model). Top k (for k=15 to 20) ranked features selected by different combinations of two and three feature selection techniques have been combined by using Boolean AND operator. Then the performance of resultant sets has been evaluated using top 10 classification algorithms. Finally the performance of all the combinations of feature sets have been compared and using Garret's Ranking Technique, ranks have been assigned to the combined feature sets (1585). The top 10 ranked combinations are mentioned in the table III. The selected reduced set of features by top 10 combinations are listed in table IV.

TABLE II. PERFORMANCE OF TOP TEN CLASSIFICATION ALGORITHMS USING 41 ATTRIBUTES

Name of Classifier	ROC Area	FPR	Accuracy	Kappa	Mean Absolute Error	Recall	Precision	Training Time	Rank
Rotation Forest	0.998	0.001	96.4	0.9053	0.0173	0.964	0.983	342.81	1
Random Tree	0.980	0.002	96.14	0.8993	0.0154	0.961	0.979	0.57	2
Random Committee	0.996	0.002	96.11	0.8982	0.0181	0.961	0.982	6.53	3
Random Forest	0.966	0.002	96.12	0.8932	0.0203	0.961	0.979	6.29	4
IBK	0.993	0.002	96.08	0.8973	0.157	0.961	0.977	0.04	5
Random Sub Space	0.999	0.002	96.07	0.8967	0.0227	0.961	0.971	18.71	6
IB1	0.979	0.002	96.08	0.8973	0.157	0.961	0.977	0.07	7
Part	0.976	0.001	95.41	0.8818	0.0188	0.954	0.978	20.16	8
Jrip	0.994	0.002	95.24	0.8778	0.0193	0.952	0.978	69.25	9
NB Tree	0.997	0.003	95.26	0.8774	0.018	0.953	0.956	195.06	10

From the table II. It can be observed that Rotation Forest Classification Algorithm perormed best in all aspects showing highest accuracy of 96.4 % but has the drawback of having highest training time. It needs to be reduced. Now a combinational aproach for feature selection is being proposed below in the table III. In which the combination of Symmetric and Gain Ratio with IBk classification performs best in all aspects with highest accuracy of 98.5 % and minimum training time of 0.03 seconds with ten features only. So combinational aproach is best for Feature selection.

TABLE III. TOP TEN COMBINATIONS OF FEATURE SELECTION TECHNIQUES AND CLASSIFIER

Name of Technique	Classifier	Training Time	ROC	Accuracy	FPR	Recall	Precision	Error	Kappa	Rank
Symmetric+ Gain Ratio (15)	IBK	0.03 seconds	0.996	98.5	0.001	0.991	0.992	0.0001	0.992	**1**
One R+ Symmetric (17)	Random Committee	7.83 seconds	0.996	97.34	0.001	0.97	0.985	0.0105	0.982	**2**
One R+Relief (21)	IB1	0.05 seconds	0.986	97	0.003	0.974	0.98	0.0104	0.983	**3**
Symmetric+ info gain (20)	IBK	0.04 seconds	0.982	96.2	0.005	0.973	0.98	0.0108	0.983	**3**
oner+symm+gain ratio (17)	Random Committee	4.5 seconds	0.982	95.4	0.004	0.967	0.987	0.0197	0.992	**4**
Symmetric+gain ratio (15)	IB1	0.03 seconds	0.986	96	0.002	0.974	0.977	0.0105	0.994	**4**
Symmetric+info gain (18)	IBK	0.03 seconds	0.971	95.2	0.004	0.973	0.975	0.0108	0.987	**5**
One R+Relief (17)	IBK	0.05 seconds	0.97	95	0.005	0.965	0.96	0.0011	0.992	**6**
One R+ Symmetric (21)	IB1	0.04 seconds	0.986	94.6	0.003	0.98	0.972	0.0105	0.985	**7**
Symmetric+ Gain Ratio (24)	IBK	0.03 seconds	0.992	94	0.004	0.949	0.95	0.0023	0.978	**7**
One R+Relief (20)	IB1	0.05 seconds	0.967	94	0.002	0.945	0.948	0.0235	0.986	**8**
Gain ratio+info gain (15)	IBK	0.03 seconds	0.971	94	0.002	0.962	0.958	0.0029	0.972	**8**
Symmetric+ Info gain (17)	IB1	0.04 seconds	0.992	93	0.002	0.958	0.959	0.0023	0.962	**9**
One R+Relief (19)	Random Tree	0.75 seconds	0.991	92	0.002	0.962	0.978	0.023	0.968	**10**

TABLE IV. LIST OF FEATURES FOR TOP TEN COMBINATIONS

Rank	Name of Technique	Features Selected
1	Symmetric+ Gain Ratio (15)	23,6,2,37,3,25,12,4,36,5 (10)
2	One R+ Symmetric (17)	23,32,5,24,33,34,29,6,3,36,2,12,37,25,35,4 (16)
3	One R+Relief (21)	23,32,5,24,33,34,29,6,3,36,2,12,25,4,26,37,30,39,35,40,38 (21)
3	Symmetric+ info gain (20)	23,32,5,33,24,34,29,3,36,6,2,12,37,4,35,38,25,31,39 (19)
4	oner+symm+gain ratio (17)	23,5,6,2,3,36,12,37,25,4,33 (11)
4	Symmetric+gain ratio (15)	23,6,2,37,3,25,12,4,36,5 (10)
5	Symmetric+info gain (18)	23,32,5,33,24,34,29,3,36,6,2,12,37,4,35,38,25 (17)
6	One R+Relief (17)	23,32,5,24,33,34,29,6,3,36,2,12,25,4 (14)
7	One R+ Symmetric (21)	23,32,5,24,33,34,29,6,3,36,2,12,37,25,35,4,39,26,30,31 (20)
7	Symmetric+ Gain Ratio (24)	23,6,2,37,3,25,12,4,36,5,33,38,32,39,31,34,26,24 (18)
8	One R+Relief (20)	23,32,5,24,33,34,29,6,3,36,2,12,25,4,26,37,30 (17)
8	Gain ratio+info gain (15)	6,2,37,3,23,12,4,36,5 (9)
9	Symmetric+ Info gain (17)	23,32,5,33,24,34,29,3,36,6,2,12,37,4,35,38 (16)
10	One R+Relief (19)	23,32,5,24,33,34,29,6,3,36,2,12,25,4,26 (15)

VIII. CONCLUSION

In this work, various feature selection techniques and classification algorithms have been reviewed. All the experiments in this work have been performed using Filter model. A combinational feature selection approach that combines the reduced feature sets resulted by two and three feature selection techniques using Boolean AND operation has been proposed. Rotation Forest is the best classification algorithm having highest rank but has large training time. User Classifier is the worst classification algorithm having lowest rank Performance of every reduced feature set varies, as it depends upon the performance of classification algorithm. Combination of Symmetric and Gain Ratio while considering top 15 attributes (reduced 10) has highest performance.

The performance using combinational approach of feature selection is much better than the performance of features set selected by individual feature selection techniques. At present this work has been focused only on Filter Model, but in future this work can be extended to evaluate another feature selection models such as Wrapper or Hybrid model.

REFERENCES

[1] R. Dash, Selection of the Best Classifier from Different Datasets Using WEKA, IJERT, Vol.2 Issue 3, March 2013.\

[2] H. Nguyen and D. Choi, Application of Data Mining to Network Intrusion Detection: Classifier Selection Model, @Springer Verlag Berlin Heidelberg, 2008.

[3] M. Panda and M. Patra, A Comparative Study of Data Mining Algorithms For Network Intrusion Detection, IEEE, First International Conference on Emerging Trends in Engineering and Technology." 2008.

[4] M. Panda and M. Patra, Ensembling Rule Based Classifiers for Detecting Network Intrusions, IEEE Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[5]B. Neethu,, Classification of Intrusion Detection Dataset using machine learning Approaches, IJECSE,2013.

[6]S. Garcia and F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons" , Journal of Machine Learning Research 9, 2008.

[7] M. Othman and T. Yau, Comparison of Different Classification Techniques using WEKA for Breast Cancer, 2012.

[8]Kdd cup 99 intrusion detection data set..Online Available:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[9] L. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, LI; 2004) [book review]." Neural Networks, IEEE Transactions on18.3 (2007): 964-964.

[10] P. Sangkatsanee, N. Wattanapongsakorn and C. Charnsripinyo. "Practicle Real- time intrusion detection using Machine learning approaches." Computer Communications (2011): 2227–2235.

[11] S. Yang, K. Chi Chang , H. Wei and C. Lin. "Feature weighting and selection for a real-time network intrusion detection system based on GA with KNN." Intelligence and Security Informatics (2008): 195-204.

[12]S. Mukherjee and N. Sharma "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Procedia Technology 4 (2012) 119 – 128

[13] Sajja, A. A, Knowledge Based Systems. Jones and Bartlett, 2012.

[14]Rich, E., Artificial Intelligence (3rd Ed.). Tata McGraw Hill, 2013.

[15] Pang-Ning, T. V., Introduction to Data Mining. Pearson. 2013.

[16] Alpaydin, E., Introduction to Machine Learning (2nd Ed.). PHI, 2010.

[17]Flach, P., Machine Learning the Art and Science of Algorithms that Make sense of Data. Cambridge, 2012.

[18] Jiawei Han, M. K., Data Mining Techniques and Concepts (3rd Ed.). Morgan Kauffman, 2013.

[19] Ian H. Witten, F. E., Practical Machine Learning Tools and Techniques (2nd ed.), 2012.

[20]V. Labatut and H. Cherifi, Evaluation of Performance Measurse for Classifiers Performance

[21] Liu H, Motoda H, Setiono R. & Zhao Z., Feature Selection: An Eve Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings Vol.4, *Publisher:*Citeseer, pages 4-13, 2010.

[22]Vege, Sri, H., Ensemble of Feature Selection Techniques for High Dimensional Data (Published Master's Thesis). Western Kentucky University, 2010.

[23] Y. Wang, F. Makedon, Application of ReliefF feature filtering algorithm to selecting informative genes for cancer classification using microarray data, Computational systems bioinformatics conference, 2004 IEEE, pages 497 – 498.

[24] S. Pulatova, Covering (rule-based) algorithms Lecture Notes in Data Mining., World Scientific publishing Co, pages 87-97, 2006.

[25]D. Ienco, R. G. Pensa, R. Meo, Context-based Distance Learning for Categorical Data clustering, LNCS 5772, Springer, Berlin, pages 83 – 94, 2009.

[26] T. Garg and S.S. Khurana, Comparison of Classification Techniques for Intrusion Detection Dataset Using WEKA, Proceedings IEEE, International Conference on Recent Advances and Innovations in Engineering, 2014.