An Ensemble Classifier Approach on Different Feature Selection Methods for Intrusion Detection

H. P. Vinutha(✉) and B. Poornima

Bapuji Institute of Engineering & Technology, Davangere, Karnataka, India
{vinuprasad.hp,poornimateju}@gmail.com

Abstract. Knowing a day's monitoring and analyzing events of network for intrusion detection system is becoming a major task. Intrusion detection system (IDS) is an essential element to detect, identify, and track the attacks. Network attacks are divided into four classes like DoS, Probe, R2L, and U2R. In this paper, ensemble techniques like AdaBoost, Bagging, and Stacking are discussed which helps to build IDS. Ensemble technique is used by combining several machine learning algorithms. Selection of features is one of the important stages in intrusion detection model. Some feature selection methods like Cfs, Chi-square, SU, Gain Ratio, Info Gain, and OneR are used in this paper with suitable search technique to select the relevant features. The selected features are applied on AdaBoost, Bagging, and Stacking with J48 as a base classifier and along with that J48 and PART are used as single classifies. Finally, results are shown that the use of AdaBoost improves the classification accuracy. Experiments and evaluation of the approaches are performed in WEKA data mining tool by using benchmark dataset NSL-KDD '99.

Keywords: Intrusion detection system · Feature selection · Ensemble techniques · WEKA · Classification accuracy

1 Introduction

Intrusion detection is a process of monitoring and analyzing event of network traffic for the signs of intrusion. Big amount of data contained on the network day by day increases the intrusions. The prevention technologies like firewall and access controls are failed to protect networks and systems from the increase of complicated attacks. Intrusion detection system (IDS) is becoming an essential element to detect, identify, and track the attacks. IDS are able to scan the network activity to recognize the attacks. There are two different intrusion detection approaches called misuse detection system and anomaly detection system. In misuse detection, the attacks are determined on the basis of pattern that are based on the known intrusions. In anomaly detection, the attacks are determined on the basis of patterns that take the deviation from normal behavior of the system. To monitor the network, IDS has an alarm system; it generates an alarm to notify that the network is under attack. It can generate four different types of attacks like True positive when legitimate attack occurs, False positive when no attack occurs, False negative when actual attack occurs, and True negative when no

V. Bhateja et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 672,
https://doi.org/10.1007/978-981-10-7512-4_44

attack occurs. So the main focus of the intrusion detection system is to increase the detection accuracy and minimize the false alarm rate.

Deploying an effective intrusion detection system is a challenging task, because dataset contains larger number of irrelevant features and redundant features. If IDS examines the entire data feature to detect intrusion, analysis becomes difficult because large number of features make it difficult to detect the suspicious behavior pattern. This reduces the learning performance and computational efficiency. So, before applying any data mining techniques like classification, clustering, association rule, and regression on the dataset, it is necessary to reduce the dimensionality of the data. A preprocessing step called feature selection is used to reduce the dimensionality. Once features are selected the classification techniques are applied on the reduced dataset to increase the performance and efficiency. Instead of using single classifiers, an ensemble classifiers are used which combine multiple classifiers to improve the accuracy and performance.

In this paper, we have discussed the different feature selection algorithms and compared the results for number of selected features and number of features removed. Then different classifiers are applied on the reduced data for all the feature selection algorithms, and the results are compared to show the best classifier. The experiment is conducted on NSL-KDD'99 dataset which was developed by Massachusetts Institute of Technology (MIT) in 1999 which is an advanced version of KDDCUP'99. The dataset contains 125973 single connection records with no redundancy with 41 features and 5 classes; they are classified as normal and 4 category of attacks: Denial of service attack (DOS), Probing attack (Probe), Remote to Local attack (R2L), and User to Root attack (U2R) [1].

Section 2 explains the related work. Section 3 discusses about feature selection algorithms. Section 4 gives the details of ensemble classifiers. Proposed method, results, and discussions are explained in Sects. 5 and 6. Section 7 concludes the paper.

2 Related Works

K Umamaheshwari et al. in this paper [2] author has proposed classification techniques on KDD-99 dataset which is a model finding process that is used for portioning the data into different classes. They have evaluated the performance of a comprehensive set of classifier algorithms Random forest, Random tree, and j48, etc. WEKA is used to compare the performance, and finally, they have concluded that Random tree algorithms produce better accuracy.

Rajender Kaur et al. [3] propose a method to deal with large amount of features which represents the whole dataset. They have done some feature selection and machine learning approaches to design the intrusion detection systems which are going to classify the network traffic data into intrusive traffic and normal traffic. Estimation is done for seven classification algorithms like Bayes Net, Naïve Bayes, J48, Random Forest, OneR, PART, and Decision Tree for tenfold cross-validation on KDD-99 dataset. It is also recommended that rule-based J48, RandomForest, and OneR classifiers are used for the detection of various attack classes.

In papers [4–7], authors have proposed different methods to deal with ensemble techniques like AdaBoost, Bagging, and Stacking. These ensemble techniques are combined with different machine learning algorithms. The experiments showed that the better results are obtained for ensemble technique combined with other algorithms than the use of single machine learning algorithms. They have used KDD-99 dataset for an experimental purpose. The preprocessing step is done on the dataset using different feature selection algorithms.

In papers [8, 9], authors have discussed various feature selection algorithms applied on dataset to select the relevant features which is used to classify the accuracy of classifiers. There are different categories of feature selection techniques like Filter method, Wrapper method, and Embedded method. The feature selection algorithms like Cfs subset, InfoGain, Gain Ratio, Filtered Attribute, Randomized hill climbing, Genetic algorithms are discussed in the paper. They have analyzed those set of algorithms with the use of different search methods, and some of the relevant features are selected by removing the irrelevant features. The best selected features are applied on different classifiers to show the improvement of performance and accuracy.

3 Feature Selection

To build an intrusion detection model, feature selection is one of the most important steps. Network data contain large number of features, but it is not good practices of using all these features. Because in the network, it is necessary to reduce the processing time to achieve the higher detection rate and accuracy. Feature selection is one of the important data preprocessing techniques in data mining. Feature selection is also known as an attributes selection method. There are three feature selection methods like Filter method, Wrapper method, and Embedded method but Filter and Wrapper methods are the commonly used methods. In Filter approach, features are selected without depending on any classifier. In Wrapper method, features are selected with the dependent on classifier [10]. Comparing the filter method, wrapper method is more time consuming because it is strongly coupled with induction algorithm which repeatedly calls the subset of features to evaluate the performance.

3.1 Importance of Feature Selection

- To reduce the size of the problem.
- Removal of irrelevant features which improves the performance of learning algorithms.
- Reduction of features reduces the storage requirement.

In this paper, we have concentrated on six different feature selection methods like Correlation Attribute Evaluation (Cfs), Chi-squared, Symmetrical Uncertainty (SU), Gain Ratio, Information Gain, and OneR. The search techniques like Best-First search and Rankers method are used with feature selection algorithms to rank the features. The ranking denotes how useful the feature which is to be classified [10].

3.2 Correlation-Based Feature Selection (Cfs)

In this algorithm, filter method is used to select the attributes. Cfs measures the individual feature by heuristic approach. Maximum value obtained for correlated and irrelevant features is avoided. Equation 1 is used to calculate the irrelevant and redundant features [11].

$$F_s = \frac{Nr_{ci}}{N + N(N-1)r_{jj}} \tag{1}$$

where N indicates the number of feature in the subset, r_{ci} says the mean feature correlation with the class, and r_{jj} means average feature inter-correlation.

3.3 Chi-Squared (X^2 Statistic)

Chi-square feature selection algorithm uses the filter method, and it calculates the need of independence between term and class for one degree of distribution freedom [12]. The expression is as Eq. 2 [12].

$$X^2_{(t,c)} = \frac{D * (PE - MQ)^2}{(P+M) * (Q+N) * (P+Q) * (M+N)}. \tag{2}$$

where D indicates the total number of documents, P says the number of documents of class C containing term t, Q means the number of documents containing t occurs without C, M indicates the number of documents class C occurs without t, N is the number of documents of others class without t [12].

3.4 Symmetrical Uncertainty (SU)

In this algorithm, set of attributes are calculated by measuring the correlation between feature and target class [13], and it is given in Eq. 3 [13].

$$SU = \frac{H(X) + H(Y) - H(X/Y)}{H(X) + H(Y)}. \tag{3}$$

where H(X) and H(Y) = entropies based on the probability associated with each feature and class value, respectively, and H(X, Y) = The joint probabilities of all combinations of values of X and Y [13].

3.5 Gain Ratio Attribute Evaluation

In this method, Gain Ratio is measured by evaluating the gain with respect to the split information. The equation is given in Eq. 4 [14].

$$\text{Gain Ratio}(A) = \text{Gain}(A)/\text{Split Info}(A) \tag{4}$$

3.6 Information Gain

In this, score is found based on how much maximum information is obtained about the classes when we use that feature. The Information Gain equation is shown in Eq. 5 [13],

$$\mathrm{IG(X) = H(Y) - H(Y|X)} \tag{5}$$

where H(Y) and H(Y|X) say the entropy of Y and the conditional entropy of Y for given X, respectively [13].

3.7 One Rule (OneR)

This is a simple classification algorithm. Classification rule is very simple and accurate in this. One level decision tree is generated by this. The rule with the smallest error rate in the training data is selected for each attribute [15].

4 Ensemble Methods

In Ensemble method, the performance of classifier is improved by combining the multiple single classifiers. Compared to single classifier, ensemble techniques are more effective and efficient. Divide and conquer approach are used in ensemble methods [16]. In this method, complex problem is divided into small subproblems which are easy to analyze and solve. Advantage of this approach is that they can get more accuracy than single algorithm. Base model is used to classify the data. In this paper, we have evaluated three different ensemble classifier techniques called Boosting, Bagging, and Stacking are used with J48 and PART classifiers [16].

4.1 Bagging

Bagging is also known as Bootstrap Aggregation. It is the simple ensemble method used to improve unstable classification problem. Variance of a predictor is reduced by this method [17]. N number of training set are created by selecting one point of the training set without the replacement of N examples. N indicates the size of original training set. Each of these datasets is used to train a different model [18].

4.2 AdaBoost

To construct a strong classifier, AdaBoost algorithm is used which is one of the most widely used Boosting techniques. The performance of individual classifiers is constructed by AdaBoost classifier [19]. It improves the performance of weak classifier by its ensemble structure. In boosting method, a set of weights is maintained across the dataset. The objects acquire more weights to classify by forcing subsequent classifier to focus on them. These methods work well by running the learning algorithm repeatedly and then combining the classifier to produce the single classifier [19].

4.3 Stacking

Stacking is the method in which we combine various classifiers to increase the efficiency. The combination of classifier is done step by step where the output of first classifier is given as an input to the second classifier. In stacking, whole dataset is divided into n number of partitions. Out of these n numbers of partitions consider two disjoint sets to use it for the first classifier. If it is S_{ij} then, i denotes number of partitions and j denotes the two disjoint set. Stacking works in two stages. Stage1 is a base learner where dataset is used on various models. A new dataset is obtained and instances of that dataset are used for prediction purpose. In stage 2, it takes the new dataset as input and gives the final output [17].

5 Proposed Methodology

Figure 1 shows the general methodology used to get the best classifier on different feature selection methods for intrusion detection system. Firstly, classify the attack types know as Normal, DoS, Probe, U2R, and R2L and save the dataset into an ARFF file format. In this proposed work, performance is analyzed by using data mining tool called WEKA3.6. The NSL-KDD'99 dataset which contains 125973 labeled connection records is used to analyze the performance. Full dataset is applied on the attribute selection algorithms called Cfs, Chi-square, SU, Information Gain, Gain Ratio, and OneR to compute the feature selection and to evaluate the classification performance on each of these feature sets. We have selected three meta-classifiers called AdaBoost, Bagging, and Stacking, and one decision tree classifier called J48 and one rule-based classifier called PART are used with full training set and tenfold cross-validation for testing purpose.

The different parameters are used to analyze the result of classification model with True positive (TP) rate, False positive (FP) rate, Kappa statistics, ROC area, Classification Accuracy. The confusion matrix summarizes the number of instances calculated normal or abnormal by the classification model.

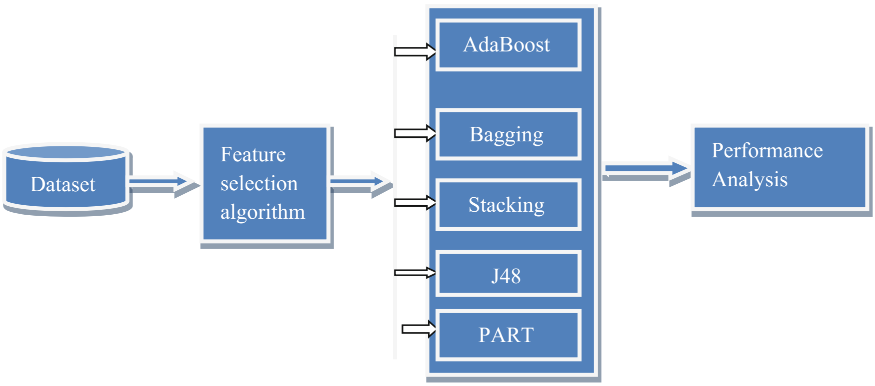

Fig. 1. Proposed methodology

6 Experimental Results and Analysis

In this experiment, we have analyzed various feature selection approaches with the help of different search methods; then the best subset of features is selected to perform on classifiers. The comparative analysis of each classifier for all the feature selection algorithms is given. The optimally selected subset of features is used on classifiers. Numbers of features removed are listed in Table 1, and they are observed to remove commonly selected features among all the approaches. They are considered separately and again to perform on classifiers.

Table 1. Number of features selected and list of features removed by different feature selection approaches

Feature selection approach	No. of features selected	List of features removed
Best-first + CfsSubsetEvl	11	1, 2, 7, 8, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 31, 32, 33, 34, 35, 36, 38, 40, 41, 42
Ranker + Chi-square	28	22, 14, 17, 13, 11, 18, 8, 16, 9, 19, 15, 7, 21, 20
Rankers + SU	27	22, 8, 13, 16, 17, 14, 19, 11, 15, 9, 7, 21, 20
Ranker + Gain Ratio	33	13, 40, 16, 19, 15, 24, 21, 7, 20
Ranker + Info Gain	30	13, 16, 17, 14, 11, 19, 18, 15, 9, 7, 21, 20
Ranker + OneR	30	7, 14, 20, 22, 19, 21, 9, 15, 18, 16, 17, 11

Number of features selected is used on three ensemble classifiers and two single classifiers. The ensemble classifiers are AdaBoost, Bagging, and Stacking. The single classifiers are J48 and PART. The percentage of correctly classified values is given in the Table 2.

Table 2. Comparison of correctly classified values for different classifiers

Classifiers/Feature selection approaches	AdaBoost	Bagging	Stacking	J48	PART
Cfs subset	**99.81**	99.75	81.66	99.77	99.77
Chi-square	**99.89**	99.81	93.39	99.82	99.77
SU	**99.89**	99.82	93.71	99.82	99.77
Gain Ratio	**99.89**	99.81	93.29	99.83	99.77
Info Gain	**99.89**	99.81	92.91	99.81	99.77
OneR	**99.89**	99.81	93.10	99.59	99.77

The above result shows that AdaBoost technique performs best than all other classifiers in accuracy rate. After observing the result of individual feature selection approaches, we have considered the number of features removed from the list. In those features, some of the common features are selected which have less importance in all the techniques and are removed to see the performance. Table 3 shows the commonly selected attributes to be removed. Table 4 shows the correctly classified values for commonly selected features removed.

Table 3. Number of commonly selected features to remove

List of features selected commonly to remove	22, 14, 17, 13, 11, 18, 16, 19, 15, 7, 21, 20, 9

Table 4. Comparative result of different classifiers for commonly selected features

Classifier	TP rate	FP rate	Precision	Recall	ROC	Accuracy
AdaBoost	**0.999**	**0.001**	**0.999**	**0.999**	**1**	**99.89**
Bagging	0.998	0.001	0.998	0.998	1	99.81
Stacking	0.938	0.027	0.933	0.938	0.955	93.76
J48	0.998	0.002	0.998	0.998	0.999	99.77
PART	0.998	0.001	0.998	0.998	0.999	99.82

In the above result of different classifiers for commonly selected features shows that AdaBoost classifier is best among five selected classifiers. Table 5 shows confusion matrix of Ensemble AdaBoost algorithm with total of each class type with accuracy. Accuracy is calculated as the sum of correct classification divided by the total number of classification.

Table 5. Confusion matrix of ensemble AdaBoost algorithm

	Normal	DoS	Probe	R2L	U2R	Total	Accuracy
Normal	67306	11	10	7	9	67343	99.94
DoS	11	45913	3	0	0	45927	99.96
Probe	33	2	11631	0	1	11667	99.69
R2L	20	0	0	954	3	977	97.64
U2R	22	0	0	1	36	59	61.01
Total	67392	45926	11644	962	40		
Accuracy	99.87	99.97	99.73	99.16	90.00		

7 Conclusion and Feature Work

In this research work, we have performed a set of experiment on classifiers at benchmark NSL-KDD'99 dataset contains 41 features. In order to remove irrelevant features from the larger dataset, the Cfs, Chi-square, SU, Gain Ratio, Info Gain, and

OneR feature selection algorithms are used. The performance of three Ensemble classifiers like AdaBoost, Bagging, and Stacking and two single classifiers like J48 and PART is compared using classification accuracy. By considering the removed features from all the algorithms, some of the common features are selected to remove and experiment is performed on the classifiers. Empirical result of experiment shows that AdaBoost classifier gives the better result. As a feature enhancement, Ensemble classifiers can be used with some other classifier as a base learning algorithm.

References

1. Vinutha H P and Poornima B: A Survey—Comparative Study on Intrusion Detection System, IJARCCE, Vol 4, Issue 7, July 2015, ISSN 2778-1021.
2. K. Umamaheswari and S. Janakiraman: Machine Learning in Networking Intrusion Detection System, ARPN journal of Engineering and Applied Sciences, Vol 11, No 2, January 2016, ISSN 1891-6608.
3. Rajender Kaur, Monika Sachdeva and Gulshan Kumar: An Empirical Analysis of Classification Approaches for Feature Selection in Intrusion Detection, IJARCSSE, Issue 9, Vol 6, September 2016, ISSN: 2277 128X.
4. Samah Osama M Kamel, Nadia H Hegazi, Hany M Harb, Adl Y S Tag El Dein, Hala Hala M Abd El Kader: AdaBoost Ensemble Learning Technique for Optimal Feature Subset Selection, IJCNCS, Vol 4, No 1, January 2016, 1–11.
5. Annkita Patel, Risha Tiwari: Bagging Ensemble Technique for Intrusion Detection System, International Journal for Technological Research in Engineering, Issue 4, Vol 2, December 2014, ISSN 2347-2718.
6. Riyad A M and M S Irfan Ahmed: An Ensemble Classification Approach for Intrusion Detection, Vol 80, No 2, October 2013.
7. Snehalata S Dongre and Kail K Wankhade: Intrusion Detection System Using New Ensemble Boosting Approach, International Journal of Modeling and Optimization, Vol 2, No. 4, August 2012.
8. S. Vanaja and K Ramesh Kumar: Analysis of Feature Selection Algorithms on Classification: A Survey, International Journal of Computer Application, Vol 96, No 17, June 2014.
9. Theyazn H Aldhyani, Manish R Joshi: Analysis of Dimensionality Reduction in Intrusion Detection, International Journal of Computational Intelligence and Informatics, Vol. 4, No. 3, October–December 2014.
10. Sheena, Krishan Kumar, Gulshan Kumar: Analysis of Feature selection Techniques: A Data Mining Approach, International Journal of Computer Applictions, ICAET 2016, IJCA2016 (1):17–21.
11. S Vanaja and K Ramesh Kumar: Analysis of Feature Selection Algorithms on Classification: A Survey, IJCA, Vol 96, No 17, June 2014.
12. Subhajit Dey Sarkar, Saptarsi Goswami: Empirical Study on Filter based Feature Selection Methods for Text Classification, International Journal of Computer Applications (0975 – 8887), Volume 81, No. 6, November 2013.
13. Zahra Karimi, Mohammad Mansour and Ali Harpunabadi: Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods, IJCA, Vol 78, No 4, September 2013.
14. Megha Aggarwal, Amritha: Performance Analysis of Difference Feature Selection Method in Intrusion Detection: IJSTR, Vol 2, Issue 6, June 2013.

15. Krishan Kumar, Gulshan Kumar, Yogesh Kumar: Feature Selection Approach for Intrusion Detection System, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 2, No. 5, Pages: 47–53 (2013) Special Issue of ICCECT 2013.
16. Iwan Syaif, Ed Zaluska, Adam Pruge-Bennett, Gary Wills: Application of Bagging, Boosting and Stacking to Intrusion Detection, Vol 28, Issues 1–2, pp, 2012.
17. NilufarZaman D P, Gaikwad: Comparision of Stacking and SVM method for KDD dataset, International Journal of Engineering Research and General Science, Issue 3, Vol 3, part 2, May–June 2015.
18. Neeraj Bisht, Amir Ahmad and A K Pant: Analysis of Classifier Ensembles for Network Intrusion Detection Systems. CAE, Vol 6, No 7, February 2017.
19. Jasmina D. Novakovic and Alempije Velijovic: AdaBoost as classifier Ensemble in classification problems, INFOTEH-JAHORINA Vol. 13, March 2014.