FEATURE SELECTION USING RANDOM FOREST IN INTRUSION DETECTION SYSTEM

Sneh Lata Pundir and Amrita
Department of CSE, Sharda University, Greater Noida, India

ABSTRACT

Intrusions are very common problem in the present scenario. To get rid of the intrusions we have the intrusion detection systems but we need to maintain the performance of the intrusion detection systems. For the intrusion detection system to be fast and effective an optimal feature subset should be made by using feature selection approaches. Therefore we proposed the approach of feature selection using random forest to improve the performance of intrusion detection systems. So, we tried to find out an optimal feature subset whose performance is equal to the performance of 41 features or near to it, so that we can make the intrusion detection systems better. We carried out our experiments on kddcup'99 dataset with random forest to select important features.

KEYWORDS: *Intrusion detection, regularized random forest, kddcup'99 dataset.*

I. INTRODUCTION

With the development of internet, network security becomes an indispensable factor of computer technology. The concept of Intrusion Detection System (IDS) proposed by Denning (1987) is useful to detect, identify and track the intruders [1]. Endorf et al., (2004) proposed a methodology in which IDS can be categorized into misuse detection that uses patterns of well-known attacks to identify intrusions and anomaly detection which determines whether deviation from the established normal usage patterns can be flagged as intrusions [2]. Dokas et al., (2002) and subsequently Wu & Yen, (2009) used data mining approaches for IDS [3][4]. These papers include the approaches Decision Trees, Naive Bayes, Neural networks, Support Vector Machine (SVM) etc., but the detection performances of these methods closely rely on the huge amount and high quality of training samples.
The aim of this paper is to select important features using the random forest approach, and the performance of these features must be equal or somehow nearby to the performance of 41 features.
The rest of the paper is organized as follows: Section II discusses related work. Section III, IV and V describes feature selection, dataset, and random forest. Section VI, VII and VIII covers proposed work, experiments and results, and conclusion.

II. RELATED WORK

P Amudha and H Abdul Rauf (2011), developed a series of experiments on the KDD Cup'99 dataset for classifying the attacks and to examine the effectiveness of Correlation feature selection measure and the hybrid and ensemble classifiers. The empirical results indicate that Random Forest gives better accuracy, detection rate and false alarm rate for DoS, Probe datasets, whereas, NB Tree gives better accuracy for R2L and U2R datasets which have small training data and better detection rate and false alarm rate for R2L dataset[5].
Sang Min Lee, Dong Seong Kim, Ji Ho Kim and Jong Sou Park, (2009) have presented a new approach named Quantitative Intrusion Intensity Assessment (QIIA) to cope with the problem of binary decision result in existing anomaly detection in IDS. Their approach's advantages are summarized as three folds. QIIA is capable of (i) identifying important features with a numerical value (ii) optimizing the parameters of Random Forests (RF) algorithm (iii) enhancing intrusion detection rates with low false positive rates through intrusion intensity based on proximity *metrics*. They have validated QIIA on KDD 1999 dataset and showed that our approach is able to identify

unknown data with intrusion intensity; this helps one to enhance and develop anomaly detection in IDS [6].

Sang Min Lee, Dong Seong Kim, Ji Ho Kim and Jong Sou Park, (2010) have presented an optimal spam detection model based on RF. They performed parameters optimization and feature selection simultaneously using RF. The advantages of their approach are summarized as four folds: It is capable of (i) optimizing the parameters of RF (ii) identifying important features as numerical value (iii) determining the optimal number of selected features by using variable importance and two threshold methods (iv) detecting spam with low processing overheads and high detection rates through (i), (ii) and (iii). They have validated their optimal spam detection model on the Spam base dataset [7].

G.Prashanth, V.Prashanth, P. Jayashree, N.Srinivasan, (2008) employed random forests algorithm in NIDS to improve detection performance. To increase the rate of minority intrusion detection, they build the balanced dataset by over-sampling the minority classes and down-sampling the majority classes. Random forests can build patterns efficiently over the balanced dataset, which is much smaller than the original one. Using random forests algorithm, a model is generated using the training set, which is used to classify the test cases. The Anomaly based detector method is used to initially identify the active routers in the system. It is in the active routers that packets are checked for corruption and attacks detected using Random Forests. They plan to implement both misuse and anomaly detection. Misuse detection can reduce false positive rate and anomaly detection can detect novel intrusions. They have started the work on multiple classifier architecture whose overall performance will be higher than the performance of each classifier built by random forests [8].

III. FEATURE SELECTION

Feature Selection is the most critical step in building intrusion detection models. Feature selection can reduce both the data and computational complexity. There are currently two methods for feature selection: *the filter method* and *the wrapper method [17]*. The filter method uses measures such as information, consistency or distance measures to compute the relevance of set of features while the wrapper method uses the predictive accuracy of a classifier as a means to evaluate the "goodness" of a feature set.

Feature selection techniques provide three main benefits when constructing predictive models [18]:

- improved model interpretability,
- shorter training times,
- enhanced generalization by reducing over fitting.

IV. DATASET

The KDD CUP 1999 benchmark datasets are used to evaluate different feature selection method for IDS [16]. It consists of 4,940,000 connection records for training data set and 311,029 connection records for test data set. The training set contains 24 attacks and the test set contains 38 attacks. Since the training and test set are prohibitively large, another 10% of the KDD Cup'99 dataset is frequently used. Each connection had a label of either normal or the attack type, with exactly one specific attack type falls into one of the four attacks categories as: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack. Each connection record consisted of 41 features and are labeled in order as 1,2,3,4,5,6,7,8,9,.....,41 and falls into the four categories as shown in the table 1 below:

Category 1 (1-9): Basic features of individual TCP connections

Category 2 (10-22): Content features within a connection suggested by domain knowledge

Category 3 (23-31): Traffic features computed using a two-second time window

Category 4 (32-41): Traffic features computed using a two-second time window from destination to host.

Given below in table 1 is the list of 41 features in the KDD cup'99 dataset:

Table 1: Lists of features in the KDD cup 99

Feature #	***FeatureName***	**Feature #**	***FeatureName***	**Feature #**	***FeatureName***
1	Duration	15	Su-attempted	29	Same-srv-rate

2	Protocol-type	16	Num-root	30	Diff-srv-rate
3	Service	17	Num-file-creations	31	Srv-diff-host-rate
4	Flag	18	Num-shells	32	Dst-host-count
5	Src-bytes	19	Num-access-files	33	Dst-host-srv-count
6	Dst-bytes	20	Num-outbound-cmds	34	Dst-host-same-srv-rate
7	Land	21	Is-hot-login	35	Dst-host-diff-srv-rate
8	Wrong-fragment	22	Is-guest-login	36	Dst-host-same-src-port-rate
9	Urgent	23	Count	37	Dst-host-srv-diff-host-rate
10	Hot	24	Srv-count	38	Dst-host-serror-rate
11	Num-failed-logins	25	Serror-rate	39	Dst-host-srv-serror-rate
12	Logged-in	26	Srv-serror-rate	40	Dst-host-rerror-rate
13	Num-compromised	27	Rerror-rate	41	Dst-host-srv-rerror-rate
14	Root-shell	28	Srv-rerror-rate		

In our proposed work we carries out our experiments with kddcup'99 unlabeled 10% test dataset. In this we carried out our experiments on 2-class problem i.e. only considered attack and normal type in the dataset and keeping that in mind we carried out our experiments.

V. RANDOM FORESTS

Random Forests (RF) is a special kind of ensemble learning techniques and robust concerning the noise and the number of attributes[9].Random forests [10] are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The term came from random decision forest that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variation.

RF builds an ensemble of CART tree classifications using bagging mechanism [11]. By using bagging, each node of trees only selects a small subset of features for the split, which enables the algorithm to create classifiers for high dimensional data very quickly. This somewhat counterintuitive strategy turns out to perform very well compared to the state-of-the-art methods in classification and regression. Also, RF runs efficiently on large data sets with many features [13] and its execution speed is fast [12]. RF produces additional facilities, especially the variable importance by numerical values [9].

The main features of random forests algorithm [14] are listed as follows:

1. It is unsurpassable in accuracy among the current data mining algorithms.
2. It shows efficient performance on large data sets with many features.
3. It can give the estimate of what features are important.
4. It has no nominal data problem and does not over fit.
5. It can handle unbalanced data sets.

A tree regularization framework is proposed by Houtao Deng and George Runger, which enables many tree models to perform feature selection efficiently. The key idea of the regularization framework is to penalize selecting a new feature for splitting when its gain (e.g. information gain) is similar to the features used in previous splits. The regularization framework is applied on random forest and boosted trees here, and can be easily applied to other tree models. Experimental studies show that the regularized trees can select high-quality feature subsets with regard to both strong and weak classifiers. Because tree models can naturally deal with categorical and numerical variables, missing values, different scales between variables, interactions and nonlinearities etc., the tree regularization framework provides an effective and efficient feature selection solution for many practical problems [15].

VI. PROPOSED WORK

In this we proposed to find out a subset out of 41 features, whose performance is equal to or greater than the performance given by the 41 features. For this purpose, we used the RRF (regularized random forest) package of r-tool [19] to rank the features with the help of their significance. We applied the code for feature selection of RRF package on the kddcup'99 dataset. Due to which we get the significance for each feature of kddcup'99 dataset and we ranked the features according to their significance. After that we used the random forest classifier of weka [20] tool to classify the feature set and check their performance.

VII. EXPERIMENT AND RESULTS

For the purpose of feature selection we ranked all the features using RRF package of R-tool and then we applied the random forest classifier of weka to check their performance. We used RRF package of R-tool for ranking of features. In this, we ranked the features by their significance. RRF package is regularized random forest package of R-tool which uses the concept of random forest. We divided the kddcup'99 dataset into ten subsets each containing 2000 records. Then we applied the code for ranking of the features using RRF package over the ten subsets of kddcup'99 dataset then we get the following ranking of features as given below in table 2 (features in the table2 below are arranged in descending order of their significance):

Table 2 : Output of feature selection method of Regularized Random Forest

S.No	Feature#	Significance	S.No	Feature#	Significance	S.No	Feature#	Significance	S.No	Feature#	Significance
1	f5	83.17135	**12**	f38	75.48408	**23**	f13	72.34996	**34**	f11	69.76801
2	f16	81.51465	**13**	f33	75.02087	**24**	f23	72.22992	**35**	f22	69.64148
3	f8	80.62985	**14**	f32	74.11562	**25**	f10	72.22949	**36**	f35	68.8579
4	f28	79.98115	**15**	f3	73.87363	**26**	f21	72.06427	**37**	f1	68.47333
5	f9	78.53039	**16**	f6	73.21583	**27**	f24	71.72387	**38**	f41	67.51357
6	f27	77.75592	**17**	f18	73.17992	**28**	f20	71.63045	**39**	f25	66.05079
7	f14	77.68277	**18**	f26	73.16692	**29**	f30	71.36763	**40**	f37	63.81618
8	f12	77.34186	**19**	f4	73.06088	**30**	f31	71.33378	**41**	f36	61.02544
9	f2	77.25005	**20**	f29	72.68134	**31**	f40	70.85286			
10	f15	77.21043	**21**	f39	72.60849	**32**	f19	70.79268			
11	f17	76.5313	**22**	f7	72.39366	**33**	f34	70.43381			

When the random forest classifier of weka tool is applied over the 41 features and it then we get the following results as shown in the figure 1.

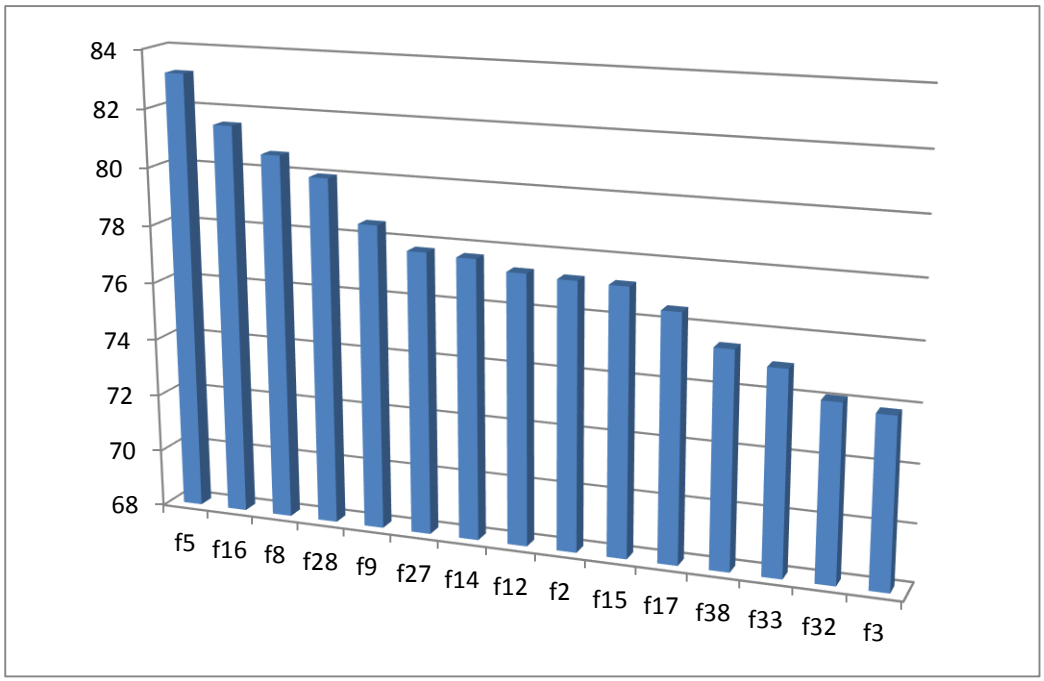

Figure 1: top 15 features selected from table 2

Below is the evaluation metrics in table 3 for the results we found out for the 41 features and its selected subsets.

Table 3 : Evaluation metrics of Random Forest for different feature set Selection method random forest

Total no of features	(41)	(3)	(5)	(10)	(15)
TP Rate	1	0.982	0.982	0.985	0.998
FP Rate	0	0.005	0.005	0.004	0
Time taken to build model (sec)	114.45	40.48	75.17	81.81	107.67
Correctly classified Instances (%)	99.97	98.19	98.19	98.51	99.80
Incorrectly Classified Instances (%)	0.03	1.81	1.81	1.49	0.20
Kappa statistic	0.9994	0.9693	0.9692	0.9748	0.9967
Mean absolute error	0.0001	0.0025	0.0025	0.0021	0.0003
Root mean squared error	0.0059	0.0356	0.0356	0.0322	0.0113
Relative absolute error (%)	0.24	4.94	4.94	4.04	0.52
Root relative squared error (%)	3.69	22.23	22.24	20.11	7.06

VIII. CONCLUSION AND FUTURE WORK

Out of all the subsets which we selected out of 41 features, the best performance is given by the subset of 15 features which is almost equal to the performance given by the set of 41 features and time taken to build the model by the subset of 15 features is less than the time taken by the set of 41 features. Hope the results given by the RRF package of r-tool and random forest classifier of weka may be helpful in future work in the field of intrusion detection and random forest.

REFERENCES

[1]. D. E. Denning, "*An Intrusion-Detection Model*", IEEE Transactions on Software Engineering, vol. SE-13,no. 2,pp.222-232, 1987.
[2]. Carl Endorf, Eugene Schultz, Jim Mellander, *Intrusion Detection & Prevention*, McGraw-Hill, 2004.
[3]. Dokas, P., Ertoz, L., Lazarevic, A., Srivastava, J., & Tan, P. N., "Data Mining for network intrusion detection", Proceeding of NGDM, pp.21–30, 2002.
[4]. Wu, S., and Yen, E., " *Data mining-based intrusion detectors*", Expert Systems with Applications, vol.36,no.3, pp.5605–5612.,2009.
[5]. P Amudha & H Abdul Rauf "Performance Analysis of Data Mining Approaches in Intrusion Detection" 978-1-61284-764-1/11/$26.00 ©2011 IEEE
[6]. Sang Min Lee, Dong Seong Kim, Ji Ho Kim & Jong Sou Park, "Quantitative Intrusion Intensity Assessment using Important Feature Selection and Proximity Metrics", 2009 15th IEEE Pacific Rim International Symposium on Dependable Computing.
[7]. Sang Min Lee, Dong Seong Kim, Ji Ho Kim & Jong Sou Park, "Spam Detection Using Feature Selection and Parameters Optimization", 2010 International Conference on Complex, Intelligent and Software Intensive Systems.
[8]. G.Prashanth, V.Prashanth, P.Jayashree & N.Srinivasan "Using Random Forests for Network-based Anomaly detection at Active routers", IEEE-International Conference on Signal processing, Communications and Networking Madras Institute of Technology, Anna University Chennai India, Jan 4-6, 2008. Pp93-96.
[9]. L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, pp. 5–32, October 2001.
[10]. http://en.wikipedia.org/wiki/Random_forest.
[11]. R. O. Duda, P. E. Hart & D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, Inc., 2001.
[12]. B.-S. Yang, X. Di & T. Han, "Random Forests Classifier for Machine Fault Diagnosis," *Journal of Mechanical Science and Technology*, Vol. 22, Jun 2008, pp. 1716-1725.
[13]. J. Zhang & M. Zulkernine, "Network Intrusion Detection using Random Forests," Proc. of the 3rd Annual Conf. on Privacy, Security and Trust (PST 2005), Oct 2005.
[14]. M. Sahami, S. Dumais, D. Heckerman & E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," Proc. of AAAI Workshop on Learning for Text Categorization, AAAI Techincal Report WS-98-05, Jul 1998.
[15]. http://arxiv.org/abs/1201.1587.
[16]. Amrita & P Ahmed vol.2, issue 3, sep 2012 1-25 © tjprc pvt. Ltd. "A study of feature selection methods in intrusion detection system: a survey".
[17]. M.Bahrololum, E. Salahi & M. Khalegi, "*Machine Learning Techniques for feature reduction in Intrusion Detection Systems: A Comparison*", Proc. of IEEE Intl. Conf. on Computer Science and Convergence Information Technology, pp.1091-1095, 2009.
[18]. http://en.wikipedia.org/wiki/Feature_selection.
[19]. The R Project for Statistical Computing, http://www.r-project.org/
[20]. www.cs.waikato.ac.nz/ml/weka/

AUTHORS

Sneh Lata Pundir is pursuing her M.Tech in Computer Science and Engineering at Sharda University, Greater Noida. She received her B.Tech in Information Technology from Gautam Buddha Technical University, Lucknow.

Amrita is an Assistant Professor in Department of Computer Science and Engineering at Sharda University, Greater Noida. She received her M.Tech. in Computer Science from Banasthali Vidyapith, Rajasthan. She is currently pursuing her Ph.D. in Computer Science and engineering from Sharda University, Greater Noida (U.P.). She has more than 12 years of experience in Academics, Software Development Industry and Government Organization.