International Journal of Pure and Applied Mathematics
Volume 118 No. 19 2018, 101-114
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Intrusion Detection Using Machine Learning: A Comparison Study

Saroj Kr. Biswas[1]
[1]CSE dept., NIT Silchar, Assam, India, 788010
saroj@cse.nits.ac.in

February 1, 2018

Abstract

With the advancement of internet over years, the number of attacks over internet has also increased. A powerful Intrusion Detection System (IDS) is required to ensure the security of a network. The aim of IDS is to monitor the processes prevailing in a network and to analyze them for signs of any possible deviations. Some studies have been done in this field but a deep and exhaustive work has still not been done. This paper proposes an IDS using machine leaning for network with a good union of feature selection technique and classifier by studying the combinations of most of the popular feature selection techniques and classifiers. A set of significant features is selected from the original set of features using feature selection techniques and then the set of significant features is used to train different types of classifiers to make the IDS. Five folds cross validation is done on NSL-KDD dataset to find results. It is finally observed that K-NN classifier produces better performance than others and, among the feature selection methods, information gain ratio based feature selection method is better.

Key Words : Intrusion detection, machine learning, NSL-KDD dataset, feature selection, classifier.

1 Introduction

Currently With the large amount of data and information, Internet has number of challenges to make it stable and secure system. Though security can be ensured through updation of firewall and software, dynamic mechanisms can also be exploited. Intrusion detection system is one of dynamic mechanisms along with network analyzers and others. Intrusion detection determines specific goal of detecting attacks [8]. Intrusion detection monitors processes prevailing in a computer system or network and analyzes them to detect any deviation or any kind of abnormalities, which are violations of computer security policies [19]. There are two methods of intrusion detection: misuse and anomaly. Misuse aims to determine attack signatures in the monitored resource. Anomaly depends on knowledge of normal behavior and any deviation from this [9]. Anomaly detection has gained popularity as it became effective against new attacks. Machine Learning algorithms can also be used for anomaly detection. Machine learning algorithms are trained and then be applied on unseen input for the actual detection process [20]. There are many classification algorithms in machine learning that can be trained and used to detect attack in a network. To further enhance the performance of these classifiers and reduce the detection time feature reduction algorithms can be used.

Sannasi Ganapathy et al. [7] presented a survey on intelligent techniques for Intrusion Detection (ID) by feature selection and classification techniques, which includes many statistical and machine learning algorithms that are used as classifiers or feature selection techniques. Vinchurkar et al. [22] analyzed the NN and other machine learning approaches in designing Intrusion Detection System (IDS). Jalil et al. [10] compared the performance of machine learning algorithms in network intrusion detection and found that DT (DT) gives better accuracy compared to SVM (SVM) and NB (NB). Amor et al [5] compared two classifiers i.e. NB and DT. They also found that DT performs better than NB.

It is observed that some comparative studies have been done in this field but exhaustive study is still not done. Therefore this paper intends to design an Intrusion Detection System (IDS) for network with a good mix of feature selection technique and classifier by

studying the combinations of most of the popular feature selection techniques and classifiers. The study will give us an idea about which feature selection technique should be combined with which classifier to build accurate network intrusion detection.

2 Feature Selection and classifiers

2.1 Feature selection

Feature selection selects representative set of attributes from the set of original attributes. This representative set keeps only the relevant and important attributes, learning algorithm takes less time to learn and produces a more general classifier as it removes unnecessary and irrelevant attributes for the original set. Feature selection also facilitates data visualization and data understanding. Some of the popular feature selection techniques used in this paper is briefly presented below.

a) ***Correlation based Feature Selection method:*** CFS works with hypothesis that is "Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other" [6].

Algorithm:

1) Select the dataset for pre-processing.
2) Calculate featurefeature and feature-class correlations.
3) Search through the feature subspace and calculate feature subset based on merit.

b) **Principal Component Analysis:** Principal component analysis (PCA) determines uncorrelated attributes called principal components.

Algorithm:

1. Whole d-dimensional dataset is taken ignoring the class labels.
2. The d-dimensional mean vector is calculated.
3. Covariance matrix is found for the whole data set.

4. Eigenvectors and corresponding eigenvalues are calculated.
5. Eigenvectors by decreasing eigenvalues are sorted and l eigenvectors with the largest eigenvalues to form a $d \times l$ dimensional matrix M are chosen.
6. M is used to transform the samples onto the new subspace. Mathematically it can be written as:$y = M^T \times p$ (Where p is a $d \times 1$ dimensional representing one sample and y is the transformed $l \times l$ dimensional sample in the new subspace).

c) ***Information Gain Ratio based feature selection:*** Features selected based on only information gain is biased towards attributes having many values. Information Gain Ratio (IGR) based Feature Selection removes this drawback by taking the splitting information of an attribute into account. Splitting information of an attribute is the entropy of pattern distribution into branches. Gain ratio of attribute decreases as value of split information increases [12].

Algorithm:

1. Start with the full set_of_attributes (set containing all attributes of the dataset) and null selected_feature_set.
2. Calculate information gain ratio of each attribute.
3. Choose an attribute from the total set with the highest information gain ratio.
4. Split the dataset into sub datasets depending on the attribute values.
5. Add the attribute to selected_feature_set and remove from set_of_attributes.
6. Repeat step 2 to 5 for each of the sub-datasets with the set_of_attributes, if instance in a sub-dataset belongs to more than one class.
7. Output the selected_feature_set.

d) ***Minimum Redundancy Maximum Relevance:*** This method tries to penalize a feature's relevance based on its redundancy. The relevance of a feature set S for the class c is defined by the

average value of all mutual information values between the individual feature f_i and the class c [16]. It is shown by equation (1):

$$\textbf{Max D(S,c) ,D} = \frac{1}{|S|}\sum_{f_i \in S} I(f_i; c) \quad (1)$$

The redundancy of all features in the set S is the average value of all mutual information values between the feature f_i and the feature f_j . It is shown by equation (2):

$$\textbf{Min R(S) ,R} = \frac{1}{|S|^2}\sum_{f_i, f_j \in S} I(f_i; f_j) \quad (2)$$

The mRMR criterion is a combination of two measures given above and is defined as shown in equation (3):

$$\textbf{Max M(D,R) ,M} = \frac{1}{|S|}\sum_{f_i \in S} I(f_i; c) - \frac{1}{|S|^2}\sum_{f_i, f_j \in S} I(f_i; f_j) \quad (3)$$

Incremental search methods are used to find the near-optimal features defined by M. Suppose we already have s_{m-1}, the feature set with (m-1) features, the mth feature from the set ($U - s_{m-1}$) will optimizes the following condition (4):

$$max_{f_j \in U - s_{m-1}}(\sum I(f_j; c) - \frac{1}{m-1}\sum_{f_i \in s_{m-1}} I(f_j; f_i)) \quad (4)$$

Algorithm:

1. Select one feature f_i from the candidate pool which has maximum value of mutual information.
2. Add f_i to subset pool and remove it from candidate pool.
3. Select the next feature f_j from the candidate pool such that it maximizes condition (5).
4. Add the selected feature f_j to subset pool and remove it from candidate pool.
5. Repeat step 3 to 4 until no more features can be added to the subset.

2.2 Classifiers

A classifier is a tool which categorizes unseen patterns in suitable classes. The classifier is first given training data which it uses to construct a decision model. Then the model is given some unseen examples to classify. Some of the popular classification techniques used in this research article is briefly presented below:

a) ***Nave Bayes:*** Nave Bayes (NB) is a form of Bayes networks which are used for inference tasks ([5],[23]). It is based on Bayes probability theory.

b) ***Support Vector Machine:*** It is based on the idea of structural risk minimization which gains advantage in speed and scalability [15]. The basic idea of Support Vector Machine (SVM) is to find a decision boundary in multidimensional space which separates unseen patterns in different classes.

c) ***Decision Tree:*** Decision Tree (DT) algorithms are well-known tool for classification and prediction tasks.It builds a model that predicts the output of a pattern based on different input attribute values of the pattern. The construction of DT does not require any domain knowledge or parameter setting, just the given data set is learnt and modelled [10]. It consists of three basic elements: Decision node, edges or branch and leaf node [4].

d) ***Neural Network:*** Neural Network (NN) is a connectionist approach which processes information through a connection of large number of artificial neurons [22]. It consists of hidden layers which further processes the data before passing the output to the output layer. To train the NN a pattern from training data set is given as input to the NN. The output obtained is observed, and if it is correct the next pattern is given as input. In case of error, the error is propagated till the input layer using back-propagation algorithm and weights are adjusted to obtain correct output for all the training patterns. Once trained the, newer unseen data is fed to the trained NN and the output obtained classifies to which class it belongs.

e) ***k- nearest neighbor algorithm (k-NN):*** This is the most simple among all machine learning algorithms where the output

is calculated based on k closest neighbours or k training patterns [1]. The calculation of output varies depending on the task to be performed. For example in case of classifying an unknown pattern, the pattern is assigned as the class which appears frequently among the k nearest training patterns.

3 Intrusion Detection System (IDS)

Fig. 1.1 shows the proposed IDS model. Feature selection method selects the significant features for the classification. Dataset with only the significant features enhances the acceptability of the model with better accuracy. The set of significant features is then used to be trained the classifiers. After training the classifier, using the same set of features the test dataset is tested to detect whether each single instance is a normal data or an attack data.

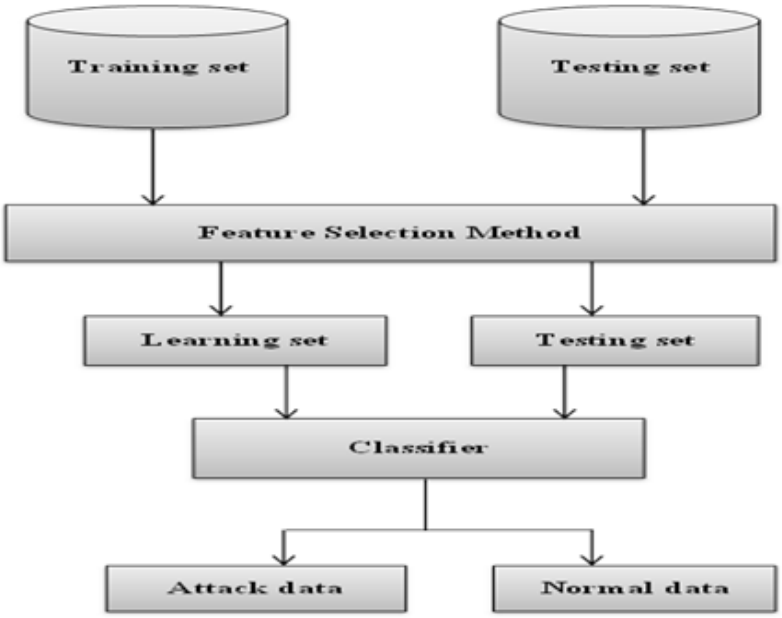

Fig.1.1. Intrusion Detection

4 Experimental Results And Discussion

NSL-KDD dataset is used for the experiments. It is a modified version of the KDD 99 data set. The modified version has some problems associated with it; some of them are discussed by McHugh [13]. But a lot analysis is done with this dataset to develop an efficient intrusion detection system due to absence of publicly available data sets for IDSs [21]. However, the size of NSL-KDD dataset is very large and has approximately about 100,000 rows of data and more than 40 columns. Large datasets are difficult to work as they increase the computational cost. Therefore, the dataset is reduced to meet the requirement.

10,000 data are selected for the experiments and are divided into 5 folds for cross validation. Four feature selection techniques and five classifiers suitable for intrusion detection ([2]-[3], [14]) are taken for the experiment. All the experiments are carried out in WEKA and effectiveness of all the classifiers in classifying the NSL-KDD data set is studied. The performances of different combinations of feature selection algorithms and classifiers in detecting normal and attack class are shown in the table 1.1.

Table 1.1. Classification accuracy of five classifiers with the four feature selection techniques

Feature Selection technique	Classifier	Accuracy (in %)
CFS	NB	82.66
	SVM	76.61
	DT	98.99
	NN	83.8
	k-NN	97.65
IGR	NB	90.29
	SVM	94.39
	DT	97.83
	NN	97.7
	k-NN	99.07
PCA	NB	89.91
	SVM	96.78
	DT	98.95
	NN	97.5
	k-NN	98.87
Minimum redundancy maximum relevance feature selection	NB	87.56
	SVM	88.93
	DT	98.78
	NN	94.6
	k-NN	98.05

The following observations are made from the results of table 1.1.

i. k-NN with IGR feature selection method produces highest performance among all the combinations. k-NN with CFS method produces lower performance than k-NN with other feature selection methods.

ii. DT with CFS method produces higher accuracy than DT with other feature selection methods and is the second highest performance among all the combinations. DT with IGR feature selection method produces lower than DT with other feature selection methods.

iii. NN with IGR feature selection method produces higher accuracy than NN with other feature selection methods. NN with CFS method produces lower than NN with other feature selection methods.

iv. SVM with PCA feature selection method produces higher accuracy than SVM with other feature selection methods. SVM with CFS method produces lower than SVM with other feature selection methods.

v. NB with IGR feature selection method produces higher accuracy than NB with other feature selection methods. NB with CFS method produces lower than NB with other feature selection methods.

It can be concluded from the observations that k-NN classifier produces better performance than others and, among the feature selection methods IGR feature selection method is better than others and CFS method is inferior to others.

5 Conclusion

This paper proposes an IDS model which compares performances of different combinations. A subset of significant features is selected using feature selection algorithms and then the set of significant features is used to train different types of classifiers. CFS, IGR,

PCA, and minimum redundancy maximum- relevance feature selection techniques, and k-NN, DT, NN, SVM and NB classifiers are used in this paper.

The paper uses different mix of feature selection algorithms and classifiers because each of the classifiers and the feature selection algorithms has advantage as well as disadvantage. It is difficult to choose one over another to implement an intrusion detection system. Further, the experimental results show that machine learning can be used in intrusion detection because all the combinations produce significant accuracy. k-NN classifier produces better performance than others and, among the feature selection methods, IGR feature selection method is better than others and CFS is inferior to others. Highest accuracy obtained in all the combinations is for IGR feature selection with k-NN. Therefore, from this study it can be concluded that the combination of IGR feature selection and k-NN can be used to design an effective intrusion detection system.

References

[1] Altman, N. S: *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, vol. 46, issue 3, pp. 175-185,(1992).

[2] Amor, N. B., Benferhat, S., and Elouedi, Z.: *NB vs DTs in Intrusion Detection Systems.* proceedings of ACM Symposium on Applied Computing, pp. 420-424,(2004).

[3] Balakrishnan, S., and Kannan, V.K.:*Intrusion Detection System Using Feature Selection and Classification Technique.* International Journal of Computer Science and Application (IJCSA), vol. 3, issue 4, (2014).

[4] Ben Amor, N., Benferhat, S., and Elouedi, Z.:*NB vs DTs in Intrusion Detection Systems.* ACM Symposium on Applied Computing, pp. 420-424, (2004).

[5] Chen, J., Huang, H., Tian, S., and Qu, Y.: *Feature selection for text classification with Nave Bayes. Expert Systems with Applications,* vol. 36, issue 3, pp. 54325435,(2009).

[6] Chen, Y., Li, Y., Cheng, X., and Guo, L.:.*Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System.* Information Security and Cryptology, Lecture notes in Computer science, 4318, pp. 153-167, (2006).

[7] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., and Kannan, A.:*Intelligent feature selection and classification techniques for intrusion detection in networks: a survey.* Journal on Wireless Communications and Networking, pp. 1-16, (2013).

[8] Gne Kayack, H., and Zincir-Heywood, N.: *Analysis of Three Intrusion Detection System Benchmark Datasets Using Machine Learning Algorithms.* Proceedings of IEEE international conference on Intelligence and Security Informatics, pp. 362-367, 2005.

[9] Gnes Kayack, H., Nur Zincir-Heywood, A., and Heywood, M. I.: *Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets.* Third Annual Conference on Privacy, Security and Trust, (2005).

[10] Jalill, K. A., Kamarudin, M. H., and Masrek, M.N.: *Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion.* International Conference on Networking and Information Technology, pp. 221-226, (2010).

[11] Jeong, D.H., Ziemkiewicz, C., Ribarsky, W., and Chang, R.: *Understanding Principal Component Analysis Using a Visual Analytics Tool.* In: Technical Report. Charlotte, Charlotte Visualization Center at UNC Charlotte, USA, (2009).

[12] Karegowda, A., Manjunath, A. S., and Jayaram, M. A.:*Comparative study of attribute selection using gain ratio and correlation based feature selection.* International Journal of Information Technology and Knowledge Management, vol. 2, issue 2, pp. 271-277,(2010).

[13] McHugh, J.: *Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory.* ACM Transactions

on Information and System Security, vol. 3, issue 4,262-294, (2000).

[14] Mukkamala, S., Janoski, G., and Sung, A.:*Intrusion detection using NNs and SVMs.* IJCNN, vol. 2, pp. 1702 1707, (2002).

[15] Osareh, A., Shadgar, B.:*Intrusion Detection in Computer Networks based on Machine Learning Algorithms.* International Journal of Computer Science and Network Security, vol. 8,(November 2008).

[16] Peng, H. C., Long, F., and Ding, C.: *Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27,issue 8, pp. 1226-1238, (2005).

[17] Quinlan, J. R.: C4.5: *Programs for machine learning.* Morgan Kaufmann Publishers Inc. San Francisco, CA, USA 1993, ISBN: 1-55860-238-0.

[18] Reddy, R.R., Kavya, B., and Ramadevi, Y.:*A Survey on SVM Classifiers for Intrusion Detection.* International Journal of Computer Applications (0975 8887) ,vol. 98, issue 19, (July 2014).

[19] Scarfone, K., and Mell, P.:*Guide to Intrusion Detection and Prevention Systems (IDPS).* National Institute Of Standards and Technology. Special Publication February-2007.

[20] Sommer, R., and Paxson, V.:*Outside the Closed World: On Using Machine Learning For Network Intrusion Detection.* IEEE Symposium on Security and Privacy, pp. 305-316, (2010).

[21] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A.:*A Detailed Analysis of the KDD CUP 99 Data Set.* Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 53-58, (2009).

[22] Vinchurkar, D. P., and Reshamwala, A.: *A Review of Intrusion Detection System Using NN and Machine Learning Technique.*

International Journal of Engineering Science and Innovative Technology, vol. 1, issue 2,(November 2012).

[23] Wagh, S. K., Pachghare, V. K., and Kolhe, S. R.: *Survey on Intrusion Detection System using Machine Learning Techniques.* International Journal of Computer Applications (0975 8887), vol. 78, issue 16, (September 2013).