ISSN: 2321-7782 (Online)

Volume 3, Issue 6, June 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A study on Performance analysis of various feature selection techniques in intrusion detection system

Harvinder Singh[1]
Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology
Hisar - India

Prof. Dharminder Kumar[2]
Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology
Hisar - India

Abstract: In these days, recognition of security threats, generally known as intrusion, has turned into a really crucial and critical problem in system, information and data security. Thus, an intrusion detection system (IDS) has turned into a really necessary portion in network or computer security. Elimination of such intrusions totally is dependent upon on detection convenience of Intrusion Detection System (IDS). As system rate becomes faster, there is an arise importance of IDS to be light with higher detection rates. Thus, several feature selection approaches/methods are planned in the literature. There are three vast kinds of strategies for choosing excellent feature subset as filter, wrapper and hybrid approach. The goal with this paper is presenting a review of numerous feature selection techniques for IDS on KDD CUP'99 bench mark dataset predicated on these three types and various evaluation criteria.

Keywords: Feature selection, intrusion detection systems, filter method, wrapper method and hybrid method.

I. Introduction

There is an increase in growth of computer network; there are various government and private organisation which store their important data on the network. So it is a challenging issue in network and information security, and detection of security threats, commonly referred to as intrusion .As it, the intrusion becomes very important issue in network, data and information security. These security attacks may cause serious problem for data and networks. Therefore, Intrusion Detection System (IDS) becomes an essential portion of every computer or network system. Intrusion detection (ID) is a mechanism that provides security for both computers and networks. In this paper we will discuss different methodologies of feature selection in IDSs.

[A] Intrusion Detection Systems:

Intrusion Detection Systems (IDS) have grown to be essential and popular used methods for ensuring network security. The breaking of security of a computer system and causing it to enter into an inferior state is known as an intrusion the intruding activity or unauthorized access is detected by an intrusion detection .but sometimes the intruder may become undetected .IDS monitors the computer or network traffic. The malicious activities are detected by the IDS and it alerts the system or network administrator against malicious attacks. An attack on a network or system is detected by the intrusion detection system. [21].IDS divide into two parts misuse detection and anomaly detection, A misuse detection system contain the previous data of the attack .if any data match with the misuse detection data then it is classified as attack, and the normal activity is referred by anomaly, Furthermore, the IDS classified into two types, Network IDS (NIDS) and Host based IDS (HIDS) system. The network are analyse by the NIDS, A single host is analysed by the HIDS.

[B] FEATURE SELECTION

Real time intrusion discovery is not possible as a consequence of plenty of data flowing upon the Internet. Feature selection can help to eliminate the particular computation as well as model complexity. Research about feature selection began at the beginning of 1960's [11]. Feature selection is actually a method of choosing a subset associated with of relevant features by eliminating a lot of unnecessary and repetitive features [12] from the data for making effective Learning models [13].

Procedure for Feature Selection:

Feature selection procedures require four basic stages in a simple feature selection method [13] shown in Figure 1.

(1) Generation procedure in order to generate the upcoming candidate subset

(2) Evaluation function so that it can evaluate the subset

(3) Stopping criterion to decide when to stop

(4) Validation procedure used for validates the subset.

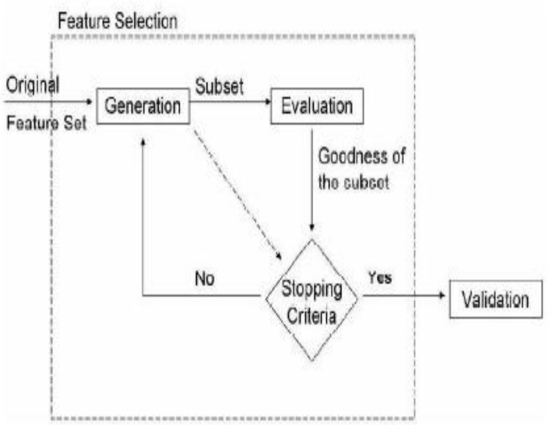

Figure 1

Techniques for Feature Selection:

Blum and Langley [12] divide the feature selection techniques into three types named filter, wrapper and hybrid (embedded) Technique.

(a)Filter: Filter technique [14] works by using additional learning algorithm in order to measure the overall performance of selected features.

(b)Wrapper: It Wrap around the learning algorithm. It works by using one pre-specified classifier to observe features .A search algorithm is used by it. The evaluation of various feature set is performed on the basic of classification performance. Which have the best performance are selected for further. Wrapper method is expensive than the filter method [15]

(c)Hybrid: The hybrid technique [15] [16] is the combination of both wrapper and filter approach .it can be used to attain most effective performance having a specific learning algorithm.

[C] DATASET DESCRIPTION: KDD'99 cup datasets is used for intrusion detection metrics. Each TCP Connection represented by 41 features. It contains 4,940,000 connection records. Each TCP connection was labelled as "normal" or “attack” with a specific attack type; the length of each connection record is 100 bytes. Simulated attack types fall in one of the following four categories [20]:

A. Denial of Service Attack (DOS): In this type of attack, the computing or memory resource becomes too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

B. User to Root Attack (U2R): In this attacker gain access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.

C. Remote to Local Attack (R2L): it occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

D. PROBE Attack: It's any attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls. Attacks could be classified based on the combination of these 41features.

II. RELATED WORK

In [2014] A relevant feature selection model was proposed to select the best features set that could be used to design a lightweight intrusion detection system. Seven different feature evaluation methods were used to select and rank relevant features. The proposed model has four different stages, Data Pre-Processing, Best Classifier Selection, Feature Reduction, and Best Features Selection. Redundant records existed in the train dataset that bias learning algorithm to the classes with large repeated records are deleted. Out of the reduced training dataset four class-based datasets have been constructed: DOS, PROBE, R2L, U2R each of these four datasets contains the attack type records + the NORMAL class records. The results indicate that certain features have no relevance or contribution to detect any intrusion attack type. Some features are important to detect all attack types, and certain features are important to detect certain attack types. A set of best 11- features have been proposed and tested against the full 41- features set. . This model is not only able to yield high detection rates but also to speed up the detection process [19].

In [2014] A model was proposed by combining an Optimal Feature Selection (OFS) algorithm and two classification techniques for securing the system. It takes more time for detecting and classifying the records using all the 41 features of the KDD’99 cup data set. The proposed feature selection algorithm selects only the important features that help in reducing the time taken for detecting and classifying the records. Further the rule based classifier and SVM help achieve a greater accuracy. The main advantage of the proposed IDS is that it reduces the false positive rates and also reduces the computation time [10].

In [2013] The performance Analysis of various classifiers is performed after reducing a number of features .They analyzes the performance of various classifier And feature selection techniques in terms of accuracy, number of features, time taken to build model, true positive rate and false positive rate are taken for the study. On the basis of these parameters comparative study is carried put. In terms of reduction of computational time, filtered subset evaluation performs the best out of these techniques. This technique reduces 82.93 % features and gives acceptable accuracy. The accuracy decrease as a result of reduction of features i.e.0.91% in Naive bayes, 0.54 % in J48, and 0.56 in PART classifier.CFS subset evaluation comes out to be the second best for the network traffic dataset. This technique reduces features by 75.61 %. [1]

In [2013] A hybrid model was proposed for feature selection and intrusion detection. They proposed a model based on PCNN and Gaussian Support Vector Machines, to intrusion detection. They found that GSVM can provide good generalization ability and effectively detect intrusions. Moreover, the modified algorithms proposed in this paper outperform conventional SVM, fuzzy-GA in terms of precision and recall. The accuracy of altered algorithms can be increase due to feature reduction of PCNN, and reduces feature sub set increase the accuracy of classification. In their experiments the PCNN-SVM can detect known attack types with high accuracy and low false positive rate which is less than 1%.[2]

In [2013] A Fuzzy Genetic algorithm was presented for intrusion detection system. The variety of features can impact on the speed and memory consumption of detection system where the reduced number of features will have faster speed and uses less memory consumption than with high number of features. Additionally, The Fuzzy Genetic Algorithm is rule-based which does not require high computation time. The Fuzzy Genetic Algorithm can identify known attack types with high accuracy and low false positive rate that is less than 1%. Moreover, the Fuzzy Genetic algorithm approach can efficiently identify new/unknown attack types with high accuracy [6]

In [2011] The problem of classification algorithm was analysed with low detection speed and low detection rate in high dimensional network data intrusion detection, A method is proposed that is based on GQPSO attribute reduction. It realizes optimal selection of network intrusion features by discarding independent and redundancy attributes. Experimental results show that classification detecting rate and detecting speed of GQPSO algorithm is higher than those of PSO and QPSO algorithms [3].

In [2011] A research work was performed which shows that Probabilistic Neural Networks provides better accuracy over Feed Forward Neural Network and Radial Basis Neural Network. To enhance the results the feature reduction methods are applied. The Principal Component Analysis is applied to the KDD CUP 1999 dataset to reduce its features and implemented using MAT LAB software. PCA selects 13 features from 41 feature data set. The reduced features are used as input to different classifiers and the results are compared. The outcomes show the efficiency with 13 features is comparable to the 41 features, with reduced training and testing set. Comparing these three classifiers PNN gives better efficiency than FFNN and RBN. The KDD Cup 1999 reduced dataset obtained with PCA shows promising results. Hence, it is proposed to consider PCA for further research [17].

In [2010], A new search method is proposed to get the globally optimal subset of relevant features by means of the CFS measure. CFS is transformed into optimization problem into polynomial mixed $0-1$ fractional programming ($P01FP$) problem. From P01FP problem, they applied their improved Chang's method to get mixed $0-1$ linear programming ($M01LP$) problem with linear dependence of the number of constraints and variables on the number of features in the full set. Branch-and-bound algorithm is used in order to solve that $M01LP$. The results showed that new approach outperforms the best-first-CFS and genetic algorithm CFS methods by removing much more redundant features and still keeping the classification accuracies or even getting better performances[5].

In [2010] A hybrid machine learning intrusion detection model was presented ,which was based on triangle area support vector machine (TASVM) .information gain was calculated for each attack class, the ten most relevant features were selected and the remaining was removed. The linearly scaling method was adopted to reprocess data for unifying their ranges. Then k-means clustering algorithm was used on the selected subset consisting of ten selected features and one label feature to produce five clustering centroids. Then chose two centroids randomly and one data point to form ten triangles and computed these triangle areas which were used in forming a new feature vector for this data. Accordingly, we could train and test a hybrid intrusion detection model with these feature vectors in LibSVM.[9]

III. Conclusion

Feature selection has grown to be essential portion within intrusion detection. Feature selection selects some sort of subset involving relevant features; it eliminates irrelevant and repetitive features from the dataset to make robust, efficient, accurate and lightweight intrusion detection system to be certain timelines for real time a lot of feature selection methods have been proposed by researchers within intrusion detection system to handle these kinds of problems. This paper has presented to survey this fast developing field and addresses the main contribution of feature selection research proposed for intrusion detection. Existing feature selection methods are reviewed as filter, wrapper and hybrid. At this time the work has been concentrated only for Filter Model, but from now on this work can be expanded in order to evaluate another feature selection models such as Wrapper or Hybrid model.

References

1. Raman Singh , Harish Kumar and R K Singla "Analysis of Feature Selection Techniques for Network Traffic Dataset,", International Conference on Machine Intelligence Research and Advancement, IEEE, pp. 21-23 ,Dec. 2013.
2. Aditya Shrivastava et.al "A Novel Hybrid Feature Selection and Intrusion Detection Based on PCNN and Support Vector Machine", Int. J. Computer Technology & Applications, IJCTA, Vol 4 (6), pp. 922-927, Nov-Dec 2013.
3. Shangfu Gong, Xingyu Gong and Xiaoru Bi., "Feature Selection Method for Network Intrusion Based on GQPSO Attribute Reduction‖", International Conference on Multimedia Technology (ICMT), pp. 6365 – 6368. In IEEE, 2011.

4. Wenke Lee , Salvatore J. Stolfo and Kui W. Mok "A Data Mining Framework for Building Intrusion Detection Models" .

5. Hai.Nguyen, Katrin Franke and Slobodan Petrovi´c, "Improving effectiveness of intrusion detection by correlation feature selection", International conference on availability, reliability and security, IEEE, pp. 17-24, 2010.

6. P. Jongsuebsuk and N. Wattanapongsakorn, "Network Intrusion Detection with Fuzzy Genetic Algorithm for Unknown Attacks," Information Networking (ICOIN), International Conference, IEEE, 2013.

7. Srinivas Mukkamala, Andrew H. Sung, and Ajith Abraham ," Intrusion Detection Using Ensemble of Soft Computing Paradigms," Intelligent Systems Design and Applications , pp 239-248, 2003

8. A.H. Sung and S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," in Proc. SAINT, pp. 209–217.,2003

9. Pingj Tang, R Jiang and Mingwei Zhao "Feature selection and design of intrusion detection system based on k-means and triangle area support vector machine". Second International Conference future network on, pp 144 – 148, IEEE 2010

10. S Balakrishnan, Venkatalakshmi K, and Kannan A "Intrusion Detection System Using Feature Selection and Classification Technique". International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 4, November 2014.

11. Lewis, P. M., "The characteristic selection problem in recognition system", IRE Transaction on Information Theory, 8, PP 171-178, IEEE 2003

12. George H John, Ron Kohavi and Karl PEger, "Irrelevant Features and the Subset Selection Problem", Proc. of the 11th International Conference. On Machine Learning, Morgan Kaufmann Publishers, pp 121-129, 1994

13. Dash, M. and Liu H., "Feature Selection for Classification", Intelligent Data Analysis, 1(3), pp 131–56, 1997

14. Liu, H. and Yu, L., "Towards integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, 17(4), pp 491-502

15. R. Kohavi and G.H. John. "Wrappers for Feature Subset Selection", Artificial Intelligence.97 (1-2), pp 273-324

16. E P Xing, M I. Jordan and R M. Karpet (2001). "Feature Selection for High Dimensional Genomic Microarray Data". Proc. 15th Int. l Conf. Machine Learning, pp 601-608.

17. S. Devaraju and Dr. S. Ramakrishnan, "Performance analysis of intrusion detection system is using various neural network classifiers", International conference on recent trends in information technology (ICRTIT), 2011.

18. John, G.H, R Kohavi and k Pfleger (1994). "Irrelevant Features and the Subset Selection Problem." Proc. of the 11th Int. Conf. on Machine Learning, Morgan Kaufmann Publishers, pp 121-129.

19. A I. Madbouly, Amr M. Gody,Tamer and M. Barakat, "Relevant Feature Selection Model Using Data Mining for Intrusion Detection System", International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 10 - Mar 2014

20. M. Tavallaee, E. Bagheri, Wei Lu, and Ali A. Ghorbani, " A Detailed Analysis of the KDD CUP 99 Data Set " , Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA 2009)

21. James P. Anderson, "Computer Security Threat Monitoring and Surveillance," Technical report, James P. Anderson Co., Fort Washington, Pennsylvania. April 1980.