Performance Analysis Of Different Feature Selection Methods In Intrusion Detection

Megha Aggarwal, Amrita

Abstract: In today's era detection of security threats that are commonly referred to as intrusion, has become a very important and critical issue in network, data and information security. Highly confidential data of various organizations are present over the network so in order to preserve that data from unauthorized users or attackers a strong security framework is required. Intrusion detection system plays a major role in providing security to computer networks. An Intrusion detection system collects and analyzes information from different areas within a computer or a network to identify possible security threats that include threats from both outside as well as inside the organization. The Intrusion detection system deals with large amount of data whichcontains various irrelevant and redundant features resulting in increased processing time and low detection rate. Therefore feature selection plays an important role in intrusion detection. There arevarious feature selection methods proposed in literature by different authors. In this paper a comparative analysis of different feature selection methods are presented on KDDCUP'99 benchmark dataset and their performance are evaluated in terms of detection rate, root mean square error and computational time.

Index Terms: Intrusion Detection, Comparative Analysis, KDDCup99 dataset, Feature Selection

1. Introduction

During the last few years there is a dramatic increase in growth of computer networks. There are various private as well as government organizations that store valuable data over the network. This tremendous growth has posed challenging issues in network and information security, and detection of security threats, commonly referred to as intrusion, has become a very important and critical issue in network, data and information security. The security attacks can cause severe disruption to data and networks. Therefore, Intrusion Detection System (IDS) becomes an important part of every computer or network system.Intrusion detection (ID) is a mechanism that provides security for both computers and networks. The paper is organized into the following sections. Intrusion Detection Systems is reviewed in Section 2. Section 3 gives the details of the datasets used in this comparative analysis. In Section 4, different methodologies of feature selection in IDSs are discussed. Related research in the literature for feature selection methods is addressed in Section 5. Section 6 presented the results drawn from comparative analysis in tabular form. Section 7 concludes the discussion over comparative analysis.

2. Intrusion Detection System

An intrusion is an attempt to compromise the integrity, confidentiality, availability of a resource, or to bypass the security mechanisms of a computer system or network. James Anderson introduced the concept of intrusion detection in 1980 [1].It monitors computer or network traffic and identify malicious activities that alerts the system or network administrator against malicious attacks.Dorothy Denning proposed several models for IDS in 1987 [2]. Approaches of IDS based on detection are anomaly based and misuse based intrusion detection. In anomaly based intrusion detection approach [3], the system first learns the normal behavior or activity of the system or network to detect the intrusion. If the system deviates from its normal behavior then an alarm is produced. In misuse based intrusion detection approach [4], IDS monitors packets in the network and compares with stored attack patterns known as signatures. The main drawback is that there will be difference between the new threat discovered and signature being used in IDS for detecting the threat. Approaches of IDS based on location of monitoring are Network based intrusion detection system (NIDS) [5] and Host-based intrusion detection system (HIDS) [6]. NIDS detects intrusion by monitoring network traffic in terms of IP packet. HIDS are installed locally on host machines and detects intrusions by examining system calls, application logs, file system modification and other host activities made by each user on a particular machine.

3. Datasets

The KDD CUP 1999 [7] benchmark datasets are used in order to evaluate different feature selection method for Intrusion detection system. It consists of 4,940,000 connection records. 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 [8] as: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack.

Denial of Service Attack (DOS): Attacks of this type deprive the host or legitimate user from using the service or resources.

Probe Attack: These attacks automatically scan a network of computers or a DNS server to find valid IP addresses.

Remote to Local (R2L) Attack: In this type of attack an attacker who does not have an account on a victim machine gains local access to the machine and modifies the data.

User to Root (U2R) Attack: In this type of attack a local user on a machine is able to obtain privileges normally reserved for the super (root) users. Each connection record

- MEGHA AGGARWAL, AMRITA
- Department of CSE, SHARDA UNIVERSITY, Greater Noida, India
- meghacs06@gmail.com, amrita.prasad@sharda.ac.in

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 are shown inTable 1:

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.

Table 1
Distribution of intrusion types in datasets

Dataset	Normal	Probe	DOS	U2R	R2L	Total
(“kddcup.data_10_percent”)	97280	4107	391458	52	1124	494020

4. Feature Selection

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s [9]. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features [10] from the data for building an effective and efficient learning model [11].

Process of Feature Selection

Feature selection processes involve four basic steps in a typical feature selection method [11] shown in Figure 1. First is generation procedure to generate the next candidate subset; second one is an evaluationfunction to evaluate the subset and third one is a stopping criterion to decide when to stop; and a validation procedure to check whether the subset is valid.

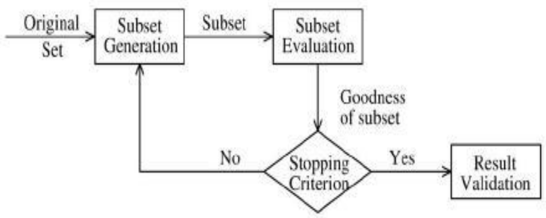

Fig1: Feature selection process [11]

Methods for Feature Selection:

Blum and Langley [12] divide the feature selection methods into three categories named filter, wrapperand hybrid (embedded) method.

Filter method: Filter method [13] uses external learning algorithm to evaluate the performance of selected features.

Wrapper method: The wrapper method [14] “Wrap around” the learning algorithm. It uses one predetermined classifier to evaluate features or feature subsets. Wrapper algorithm [15] uses a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Many feature subsets are evaluated based on classification performance and best one is selected. This method is more computationally expensive than the filter method [16] [14].

Hybrid method: The hybrid method [16] [17] combines wrapper and filter approach to achieve best possible performance with a particular learning algorithm.

5. Background

In paper [18], a feature selection approach based on Genetic Quantum Particale Swarm Optimization (GQPSO) for network intrusion detection has been proposed. In the approach, selection and variation of genetic algorithm with QPSO algorithm are combined to form GQPSO algorithm. The proposed method reduces redundant and irrelevant features. Experimental results show that classification detecting rate and detecting speed of GQPSO algorithm is higher than those of PSO and QPSO algorithms. Support VectorMachine (SVM) is used as a classifier. In paper [19], a simple Genetic Algorithm (GA) is used to evolve weights for the features and k-nearest neighbor (KNN) classifier is used as fitness function of the GA and also as classifier. One advantage of the KNN method is that it is easy to apply a weight to a feature of the data set.This weighted feature set has reduced noise present in the data and improved levels of KNN classification.Top five ranked features for each class are selected {DoS-23,29,1,11,24, R2U-24,3,12,23,36, U2R-24,6,31,41,17, Probe-2,37,30,3,6}. The result shown indicates an increase in intrusion detection accuracy. This paper [20] proposed an approach where genetic search methods along with correlation are used for feature selection and Immune system is used as a classifier. A new artificial Intelligenceparadigm called the artificial immune system (AIS) was created based on human immune system. To implement a basic Artificial Immune System, four decisions have to be made: Encoding, Similarity Measure, Selection and Mutation. Attributes are selected based on correlation based feature using genetic search. The selected features are used to train the AIS algorithm and subsequently tested. In the paper two soft computing techniques for Network Intrusion Detection System (NIDS) are used. A genetic search approach was considered for correlation based feature selection. Artificial Immune System (AIS) based classifier was used to classify the class labels over the selected features. Results obtained show recall of 99.7% for normal data. Recall of 3.5% was obtained for teardrop which had only one instance in the dataset. In paper [21], they proposed a new hybrid feature selection method – a fusion of Correlation-based Feature Selection, Support Vector Machine and Genetic Algorithm – to determine an optimal feature set. Correlation-based Feature Selection (CFS) is a filter method. It evaluates merit of the feature subset. A flow chart is given in this paper that describes the working of the proposed hybrid algorithm.The hybrid feature selection method reduced the computational resource while maintaining the detection and false positive rate within tolerable range. The proposed algorithm also reduces the training time and testing time. Faster training and testing helps to build lightweight intrusion detection system.

6. Study of feature selection methods

A number of feature selection algorithms are proposed by various authors. The aim of this work is to examine the various existing attribute selection methods in terms of detection rate andcomputational time.Out of the total 41 network traffic features, used in detecting intrusion, some features will be potential in detecting intrusions. Therefore the predominant features are extracted from the 41 features that are really effective in detecting intrusions.

Attribute evaluators [22]:

Attribute evaluator is basically used for ranking all the features according to some metric. Various attribute evaluators are available in WEKA. We used (Weka, 3.7.8) a learning machine tool in this work which includes CfsSubsetEval, ChiSquaredAttributeEval, InfoGainAttributeEval and GainRatioAttributeEval.

a. **CfsSubsetEval:** Evaluates the worth of a subset of attributes by considering the individual ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation with the other attributes are preferred.

b. **ChiSquaredAttributeEval:** Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.

c. **GainRatioAttributeEval:** Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

GainR (Class, Attribute) = (H(Class) _ H(Class|Attribute))/ H(Attribute).

d. **InfoGainAttributeEval:** Evaluates the worth of an attribute by measuring the information gain with respect to the class.

InfoGain (Class, Attribute) = H(Class) _ H(Class|Attribute).

Search Methods:

These methods search the set of all possible features in order tofind the best set of features. Four search methods which includes BestFirst, GeneticSearch, GreedyStepwise and Ranker available in weka are used in this work for comparison purpose.

a. **BestFirst:**This searches the space of attribute subsets by greedyhillclimbing augmented with a backtracking facility. Settingthe number of consecutive non-improving nodes allowed controlsthe level of backtracking done. Best first may start withthe empty set of attributes and search forward, or start withthe full set of attributes and search backward, or start at anypoint and search in both directions (by considering all possiblesingle attribute additions and deletions at a given point).

b. **GeneticSearch:**It performs a search using the simple geneticalgorithm.

c. **GreedyStepwise:** It performs a greedy forward or backwardsearch through the space of attribute subsets. May start withno/all attributes or from an arbitrary point in the space. Stopswhen the addition/deletion of any remaining attributes resultsin a decrease in evaluation. Can also produce a ranked list ofattributes by traversing the space from one side to the otherand recording the order that attributes are selected.

d. **Ranker:** It ranks attributes by their individual evaluations. Use inconjunction with attribute evaluators (Chisquare, GainRatio,InfoGainetc).

7. Results and Discussions

We used Weka (3.7.8) a learning machine tool to draw the comparative analysis. In this paper different combinationof feature selection methods are tried and they include BestFirst + CfsSubsetEval, GeneticSearch + CfsSubsetEval, GreedyStepwise + CfsSubsetEval, Ranker + ChiSquaredAttributeEval, Ranker + InfoGainAttributeEval and Ranker + GainRatioAttributeEval. The details of the combinations and the features selected by each combination and their visualization is described in Table 1, 2, 3, Figs. 2 and 3.

Table 2
List of features selected by different feature selection methods

S.No	Feature Selection Method	Number of selected features	Selected Features
1.	Bestfirst+CFSSubsetEval	11	2,3,4,5,6,7,8,14,23,30,36
2.	GeneticSearch+CFSSubsetEval	17	2,3,5,6,7,8,10,23,24,28,29, 33,35,36,37,38,39
3.	GreedyStepwise+CFSSubsetEval	11	2,3,4,5,6,7,8,14,23,30,36
4.	Ranker+InfoGainAttributeEval	25	2,3,4,5,6,12,23,24,25,26, 27,28,29,30,31,32,33,34, 35,36,37,38,39,40,41
5.	Ranker+GainRatioAttributeEval	25	2,3,4,5,6,7,8,10,11,12,13,14, 22,23,25,26,29,30,33,34,35 36,37,38,39
6.	Ranker+ChiSquaredAttributeEval	25	2,3,4,5,6,7,8,10,11,13,23,24, 25,26,27,29,30,33,34,35,36, 37,38,39,40

Table 3
Evaluation of different feature selection methods based on Naive Bayes

		Evaluation Criteria			
S.No	**Feature Selection Method**	**Detection Rate**	**Time taken to build model**	**Time taken to test model**	**Root Mean Square Error**
1.	Bestfirst+CFSSubsetEval	91.5749%	1.31s	42.51s	0.074
2.	GeneticSearch+CFSSubsetEval	95.9963%	2.32s	59.42s	0.0574
3.	GreedyStepwise+CFSSubsetEval	91.5749%	1.31s	42.51s	0.074
4.	Ranker+InfoGainAttributeEval	99.5939%	0.28s	11.22s	0.0172
5.	Ranker+GainRatioAttributeEval	99.6118%	0.33s	11.51s	0.0169
6.	Ranker+ChiSquaredAttributeEval	99.5962%	0.3s	11.32s	0.0168

Table4
Evaluation of different feature selection methods based on C4.5 (J48)

		Evaluation Criteria			
S.No	**Feature Selection Method**	**Detection Rate**	**Time taken to build model**	**Time taken to test model**	**Root Mean Square Error**
1.	Bestfirst+CFSSubsetEval	99.9587%	17.57s	2.88s	0.0057
2.	GeneticSearch+CFSSubsetEval	99.9779%	34.7s	3.57s	0.0042
3.	GreedyStepwise+CFSSubsetEval	99.9587%	17.57s	2.88s	0.0057
4.	Ranker+InfoGainAttributeEval	99.9549%	4.51s	3.42s	0.006
5.	Ranker+GainRatioAttributeEval	99.9688%	8.31s	7.56s	0.0049
6.	Ranker+ChiSquaredAttributeEval	99.968%	4.81s	4.19s	0.005

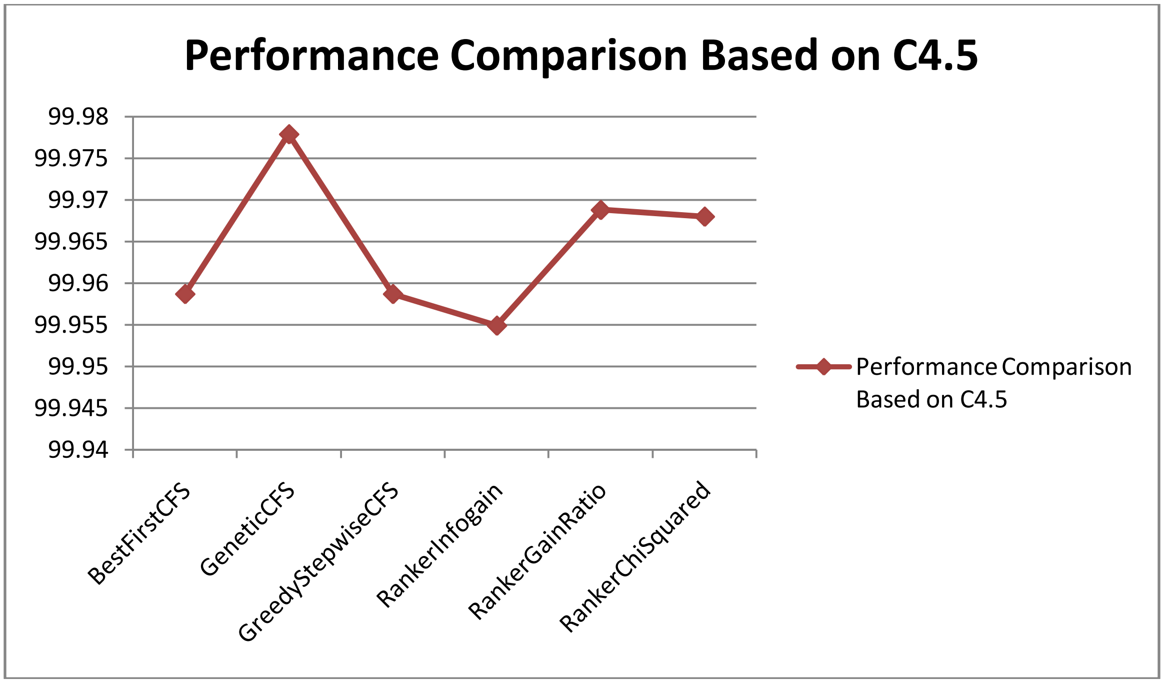

Fig2. Performance comparisons of various feature extraction algorithms based on C4.5

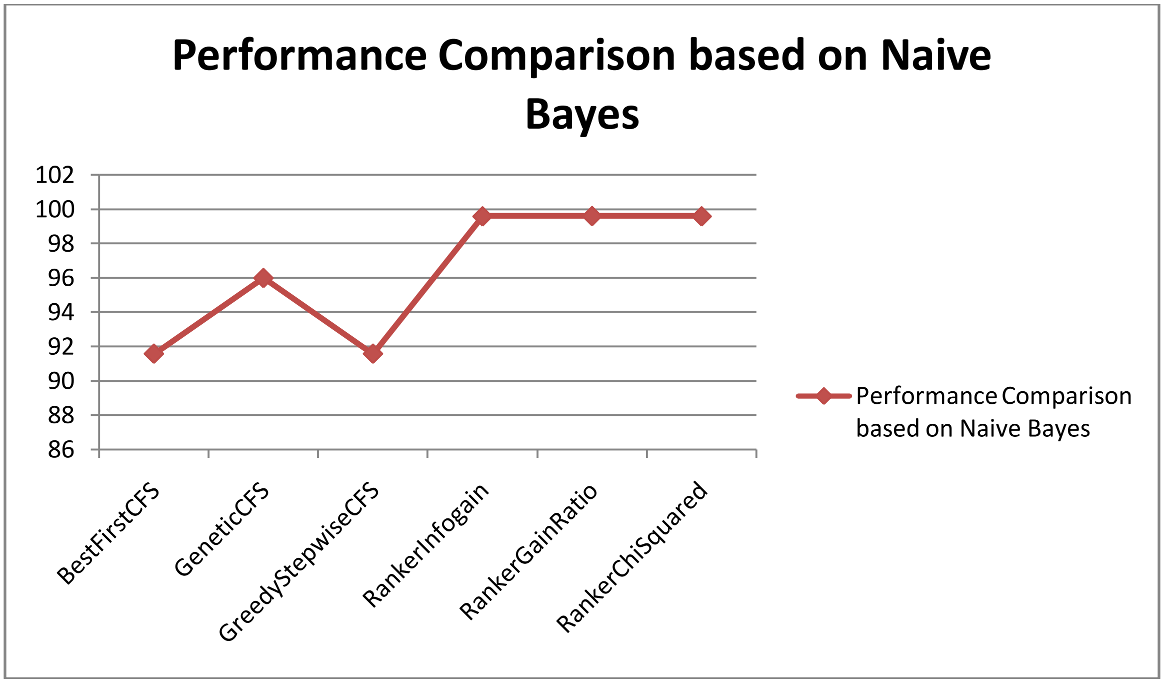

Fig3. Performance comparisons of various feature extraction algorithms based on Naive Bayes

8. Conclusion

In this paper a comparative analysis has been done on the basis of detection rate, computational time and root mean square error. In this analysis six feature selection algorithms are used and their performance is evaluated using Naïve Bayes and C4.5(J48) classifier. Features selected using Bestfirst+CFSSubsetEval and GreedyStepwise+CFSSubsetEval is same so their performance is same. GeneticSearch+CFSSubsetEval performance is good over other two and we can say CFS performs best with genetic search.InfoGain, GainRatio and Chisquared are feature selection methods that are based on ranking. So on the basis of ranking we select top 25 attributes from each of the three feature selection methods and then by doing analysis it has been observed that the performance of Ranker+GainRatioAttributeEval is good in terms of detection rate but it takes more testing and training time. Ranker+InfoGainAttributeEval takes less computational time among all the feature selection methods. In this paper two classifiers are used namely NaiveBayes and C4.5 and it has been observed that NaiveBayes takes less time to test the dataset but more time in training the set whereas C4.5 does the reverse.

References:

[1]. Anderson, James P., “Computer Security Threat Monitoring and Surveillance”, James P. Anderson Co., Fort Washington, Pa., 1980.

[2]. Denning, D. E. (1987), “An intrusion detection model. IEEE Transaction on SoftwareEngineering”, Software Engineering 13(2), 222-232.

[3]. Denning, D. E. (1987). An intrusion detection model. IEEE Transaction on Software Engineering, Software Engineering 13(2), 222-232.

[4]. Wu, S.X. &Banzhaf, W. (2010). The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing Journal, 10, 1–35.

[5]. Lazarevic, A., Ertoz, L., Kumar V., Ozgur A. &Srivastava J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In Proc. of the SIAM Conference on Data Mining.

[6]. Kumar, S. &Spafford, E. H. (1994). A pattern matching model for misuse intrusion detection.In Proceedings of the 17th National Computer Security Conference, 11-21.

[7]. sKDD Cup 1999 Intrusion detection dataset: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[8]. Mukkamala, S. et al. (2005). Intrusion detection using an ensemble of intelligent paradigms.Journal of Network and Computer Applications, 28(2), 167–82.

[9]. Lewis, P. M. (1962). The characteristic selection problem in recognition system. IRE Transaction on Information Theory, 8, 171-178.

[10]. John, G.H. et al. (1994). Irrelevant Features and the Subset Selection Problem. Proc. of the 11thInt. Conf. on Machine Learning, Morgan Kaufmann Publishers, 121-129

[11]. Dash, M. & Liu, H. (1997). Feature Selection for Classification. Intelligent Data Analysis, 1(3), 131–56

[12]. Blum, Avrim L. & Pat Langley (1997). Selection of relevant features and examples in machinelearning. Artificial Intelligence, 97(1-2), 245–271

[13]. Liu, H. & Yu, L. (2005).Towards integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502

[14]. Das, S. (2001). Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. Proc. 18thInt'l Conf. Machine Learning, 74-81

[15]. Liu, H. & Yu, L. (2005). Towards integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502

[16]. R. Kohavi and G.H. John (1997).Wrappers for Feature Subset Selection.Artificial Intelligence.97 (1-2), 273-324

[17]. Xing, E. et al. (2001). Feature Selection for High-Dimensional Genomic Microarray Data. Proc. 15th Int'l Conf.Machine Learning, 601-608

[18]. Gong, S. (2011)," Feature Selection Method for Network Intrusion Based on GQPSO Attribute Reduction", International Conference on Multimedia Technology (ICMT), 6365 – 6368.

[19]. Nyguen, H. and Franke, K. et al.(2010)"Improving effectiveness of intrusion detection by correlation feature selection", International conference on availability, reliability and security, 17-24.

[20]. Xing, E. et al. (2001)"Feature Selection for High-Dimensional Genomic Microarray Data", Proc.15th Int'l Conf.Machine Learning, 601-608.

[21]. Sridevi,R. and Chattemvelli ,R.(2012) "Genetic algorithm and Artificial immune systems: A combinational approach for network intrusion detection ",International conference on advances in engineering, science and management (ICAESM-2012),494-498.

[22]. http://weka.sourceforge.net/doc.dev/weka/attribute Selection.

Ms. Megha Aggarwal received her B.Tech degree with honors in Computer science and engineering from UPTU university. She is pursuing M.Tech in computer science and engineering from Sharda university. Her areas of interest are computer networks and security.

Ms. 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 BanasthaliVidyapith, 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.