ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, No 5, May 2013

Fusion of Statistic, Data Mining and Genetic Algorithm for feature selection in Intrusion Detection

MeghaAggarwal& Amrita

Department of computer science and engineering, Sharda University, Greater Noida

Abstract

The security of information and data is a critical issue in a computer networked environment. In our society computer networks are used to store proprietary information and to provide services for organizations and society. So in order to secure this valuable information from unknown attacks (intrusions) need of intrusion detection system arises. There are many intrusion detection approaches focused on the issues of feature reduction as some of the features are irrelevant or redundant which results in lengthy detection process and degrading the performance of an IDS. So in order to design lightweight IDS we investigate the performance of three feature selection approaches CFS, Information Gain and Gain Ratio. In this paper we propose a fusion model by making use of the three standard algorithms and finally applying genetic algorithm that identify important reduced input features. We apply Naive Bayes classifier on the dataset for evaluating the performance of the proposed method over the standard ones. The reduced attributes shows that proposed algorithm give better performance that is efficient and effective for detecting intrusions.

Keywords- CFS, InfoGain, GainRatio, Genetic Algorithm, KDDCup99 dataset, NaiveBayes, Intrusion Detection.

1. Introduction

The rapid development of computer networks and mostly of the Internet has created many challenging issues in network and information security such as intrusions on computer and network systems. 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].These security attacks can cause severe disruption to data and networks. Therefore, Intrusion Detection system becomes an important part of every computer or network system. 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 categorized either as misuse detection or anomaly detection:

Misuse detection- Misuse intrusion detection uses well-defined patterns of the attack that exploit weaknesses in the system to identify the intrusions [3].

Anomaly detection – Anomaly detection refers to techniques that define and characterize normal behaviors of the system, any deviation from this expected normal behaviors are considered as intrusions [3].

Approaches of IDS based on location of monitoring are categorized either as Network based intrusion detection system (NIDS) [4] and host based intrusion detection system (HIDS) [5]:

Network based Intrusion detection- NIDS detects intrusion by monitoring network traffic in terms of IP packet.

Host based Intrusion detection- HIDS are installed locally on host machines and detects intrusions by examining system calls, application logs, other host activities made by each user on a particular machine.

Due to large volumes of data as well as the complex and dynamic properties of intrusion behaviors, identifying intrusion traffic from normal becomes difficult. Due to this, IDS has to meet the challenges of low detection rate and large computation. Therefore, Feature selection is a very important issue and plays a key role in intrusion detection in order to achieve maximal performance. Feature selection is the selection of that minimal cardinality feature subset of original feature set that retains the high detection accuracy as the original feature set [6]. Blum and Langley [7] divide the feature selection methods into three categories named filter, wrapper [8] and hybrid (embedded) method.

Network based Intrusion detection- NIDS detects intrusion by monitoring network traffic in terms of IP packet.

Host based Intrusion detection- HIDS are installed locally on host machines and detects intrusions by examining system calls, application logs, other host activities made by each user on a particular machine.

Due to large volumes of data as well as the complex and dynamic properties of intrusion behaviors, identifying intrusion traffic from normal becomes difficult. Due to this, IDS has to meet the challenges of low detection rate and large computation. Therefore, Feature selection is a very important issue and plays a key role in intrusion detection in order to achieve maximal performance. Feature selection is the selection of that minimal cardinality feature subset of original feature set that retains the high detection accuracy as the original feature set [6]. Blum and Langley [7] divide the feature selection methods into three categories named filter, wrapper [8] and hybrid (embedded) method.

Filter method: Filter method [9] uses external learning algorithm to evaluate the performance of selected features.

Wrapper method: The wrapper method [10] "Wrap around" the learning Algorithm. It uses one predetermined classifier to evaluate subsets of features. This method is computationally more expensive than the filter method [11] [10].

Hybrid method: The hybrid method [11] [12] combines wrapper and filter approach to achieve best possible performance with a particular learning algorithm.

The rest of this paper is organized as follows. In Section 2, we review a background of feature selection. In section 3, a review of standard feature selection methods are given. Next, the proposed feature selection algorithm is presented. In Section 5, the experimental results are reported, and an implication and future direction of this study are discussed in the final section.

2. Background

In [13], the author has 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.

In paper [14], a feature relevance analysis is performed on KDD 99 training set, which is widely used by machine learning researchers. Feature relevance is expressed in terms of information gain, which gets higher as the feature gets more discriminative. In order to get feature relevance measure for all classes in training set, information gain is calculated on binary classification, for each feature resulting in a separate information gain per class.

In [15], the author has proposed an automatic feature selection based on the filter method used in machine learning. In particular, we focus on Correlation Feature Selection (CFS). By transforming the CFS optimization problem into a polynomial mixed 0−1 fractional programming problem and by introducing additional variables in the problem transformed in such a way, they obtain a new mixed 0 –1 linear programming problem with a number of constraints and variables that is linear in the number of full set features. The mixed 0−1 linear

programming problem can then be solved by means of branch-and-bound algorithm. Their feature selection algorithm was compared experimentally with the best-first-CFS and the genetic-algorithm-CFS methods regarding the feature selection capabilities. The classification accuracy obtained after the feature selection by means of the C4.5 and the Bayes Net machines over the KDD CUP'99 IDS benchmarking data set was also tested.

In paper [16] the author has incorporated information gain (IG) method for selecting discriminative features and triangle area based SVM by combining k- means clustering algorithm and SVM as a classifier for detecting attacks.

3. Feature Selection Methods

Basically there are two types of feature selection methods [20]-

Feature Ranking:

(a) Rank features according to some criterion and selects the top K features.
(b) A threshold is needed in advance to select the top K features.

Feature Subset Evaluator:

(a) Selects the minimum subset of features that does not deteriorate learning performance.
(b) No threshold necessary.

3.1 Correlation-based Feature Selection (CFS):

CFS is basically a feature subset evaluator method of feature selection. It evaluates merit of the feature subset on the basis of hypothesis –"Good feature subsets contains features highly correlated with the class yet uncorrelated to each other [17]".With CFS as attribute evaluator and search strategy such as best first is used to search the feature subset in reasonable time. Equation 1 for calculating CFS is

$$M_s = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}}$$

Where $M_s$gives the merit of a feature subset S, k is the number of features present in the feature subset r_{cf} is average feature-class correlation and r_{ff} is average feature-feature correlation [17].

3.2 Info Gain (IG):

Info Gain is basically a feature ranking method of feature selection. This method evaluates attributes by measuring their information gain with respect to the class. Let C be a set of training set samples with theircorresponding labels. Suppose there are m classes and thetraining set contains $C_i$samples of class I and C is the totalnumber of samples in the training set [14]. Expectedinformation needed to classify a given sample is calculated by:

$$I(S_1, S_2 \ldots\ldots\ldots\ldots S_m) = -\sum_{i=1}^{m} \frac{C_i}{C} \log_2 \frac{Ci}{C} - (1)$$

A feature F with values { f_1, f_2, …, f_v} can divide the training set into v subsets { C_1, C_2, …, C_v } where $C_i$is the subset which has the value $f_j$for feature F. Furthermore let $C_j$contain C_{ij}samples of class i. Entropy of the feature F is

$$E(F) = \sum_{j=1}^{v} \frac{C_{ij} + \ldots\ldots\ldots + C_{mj}}{C} * I(C_{ij} + \ldots\ldots + C_{mj}) - (2)$$

Information gain for F can be calculated as:

$$Gain(F) = I(C_1 \ldots\ldots\ldots\ldots C_m) - E(F) - (3)$$

3.3 Gain Ratio (GR):

Gain Ratio is also a method of feature ranking for feature selection. The gain ratio is an extension of info gain, attempts to overcome the bias. Gain ratio applies normalization to info gain using a value defined as

$$SplitInfo_f(C) = -\sum_{i=1}^{v} (|C_i| / |C|) \log_2 (|C_i| / |C|)$$

The value represents the potential information generated by splitting the training dataset, C, into v partitions, corresponding to the v outcomes of a test on attribute A [18].

$$GainRatio(F) = Gain(F) / SplitInfo_f(S)$$

4. Genetic Algorithms:

Genetic algorithms are basically computerized search and optimization methods that work very parallel to the principles of natural evolution. Based on

Darwin's survival-of-the-fittest principles, GA's intelligent search procedure finds the best and fittest design solutions [22]. Potential solutions to the problem to be solved are encoded as sequences of bits, characters or numbers. The unit of encoding is called a gene, and the encoded sequence is called a chromosome. Each chromosome represents one possible solution to the problem. GA is able to select subsets of various sizes in order to determine the optimum combination and number of inputs to network. A chromosome contains the information about the solution to a problem, which it represents. Typically, it can be encoded using a binary string as follows [23]:

Chromosome 1 1101100100110110
Chromosome 2 1101111000011110

In which a bit value of 1 in the chromosome representation means that the corresponding feature is included in the specified subset, and a value of 0 indicates that the corresponding feature is not included in the subset.
The set of chromosomes during a stage of evolution are called population. An evaluation function is used to evaluate the fitness of each chromosome. During the process of evaluation crossover and mutation operator are used to simulate the natural reproduction and mutation of genes. Genetic algorithm starts with a randomly generated population, evolves through selection, crossover, and mutation. Finally, the best chromosome is picked up as the final result. This allows reducing the computational expense on the training system with near optimal results still reachable. Research [21] has shown that GA is one of the most efficient of all feature selection methods.

5. Proposed Method:

In this approach detection of intrusions will be accomplished by using a fusion of feature selection approaches. There are several existing feature selection approaches but we will use a fusion of feature selection approaches by incorporating CFS, Info Gain, Gain Ratio and finally applying genetic algorithm (GA) for intrusion detection. The proposed method is discussed below.

Step1: Select features using CFS (defined in 3.1)

$$S_{CFS} = \{ f_{cfs1}, f_{cfs2}, f_{cfs3}, f_{cfs4} ----- f_{cfsn} \},\ n << 41$$

Step2: (i) Select features using Information Gain (IG). These features are arranged on the basis of their rank (defined in 3.2)

$$S_{IG\_T} = \{ f_{IG1}, f_{IG2}, f_{IG3} ..... f_{IGn} \},\ n << 41$$

(ii) From set S_{IG_T} ,select top 30 ranked features.

$$S_{IG(30)} = \{ f_{IG1}, f_{IG2}, f_{IG3} ..... f_{IG30} \}$$

Step3: (i) Select features using Gain Ratio (GR). These features are arranged on the basis of their rank (defined in 3.3)

$$S_{GR\_T} = \{ f_{GR1}, f_{GR2}, f_{GR3} ..... f_{GRn} \},\ n << 41$$

(ii) From set S_{GR_T} ,select top 30 ranked features.

$$S_{GR(30)} = \{ f_{GR1}, f_{GR2}, f_{GR3} ..... f_{GR30} \}$$

Step4: Apply union operation on the sets obtained from steps (1), (2) and (3).

$$S_T = (S_{CFS} \bigcup S_{IG(30)} \bigcup S_{GR(30)})$$

Step5: Finally applying Genetic algorithm (GA) on the set S_T.

Step6:Evaluate the performance of the set S_T using Naïve Bayes classifier.

6. Experimental Setup:

We used WEKA 3.7.8 a machine learning tool [19], to compute the feature selection subsets for

CFS, IG, GR and the proposed algorithm and also to measure the classification performance on each of these feature sets. We have used "kddcup.data_10_percent" dataset for evaluating the performance of the proposed method. Each connection had a label of either normal or attack type and the attack type can be further classified into four categories namely DOS, probe, U2R and R2L.

1. **Denial of Service Attack (DOS):** Attacks of this type deprive the host or legitimate user from using the service or resources.
2. **Probe Attack**: These attacks automatically scan a network of computers or a DNS server to find valid IP addresses.
3. **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.
4. **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.

We have used naive bayes classifier for evaluating the performance of our proposed method.

7. Result

Basically we used three standard methods and one proposed method for feature reduction. The feature reduction is performed on 41 features and obtained 11, 30, 30 and 17 features.

Table 1: List of features selected by different feature selection methods

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

Table 2: Performance of feature reduction methods

Feature Reduction Methods	No. of attributes	Time taken to build model	Time taken to test model	Accuracy
CFS+BestFirst	11	1.31s	42.51s	91.5749 %
InfoGain +Ranker	30	0.35s	12.88s	99.6249 %
GainRatio+Ranker	30	0.3s	12.85s	99.6421 %
All Features	41	0.34s	17.21s	99.6466 %
Proposed Method	17	0.21s	8.88s	99.6563 %

The reduced feature set obtained in proposed algorithm is smallest among the standard feature selection algorithms and it performs better than other methods in terms of detection rate and computational time. The figure below shows comparative graph for classifier accuracy on the reduced features obtained by (i) CFS+bestfirst (ii) IG+Ranker (iii) GR+Ranker(iv) Proposed method.

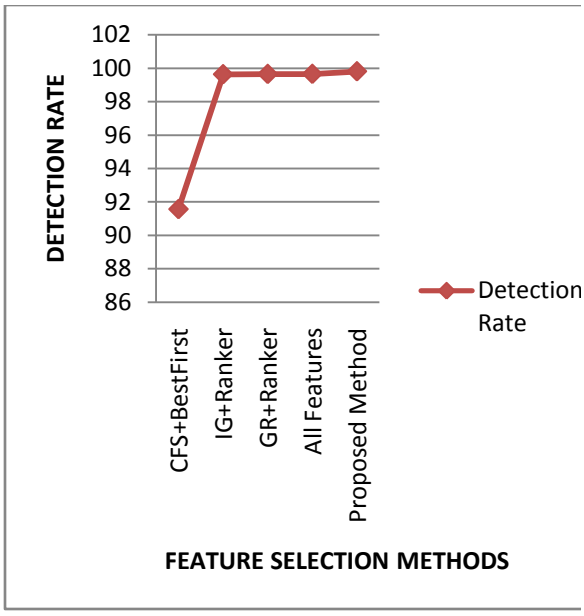

Figure 1: Detection rate

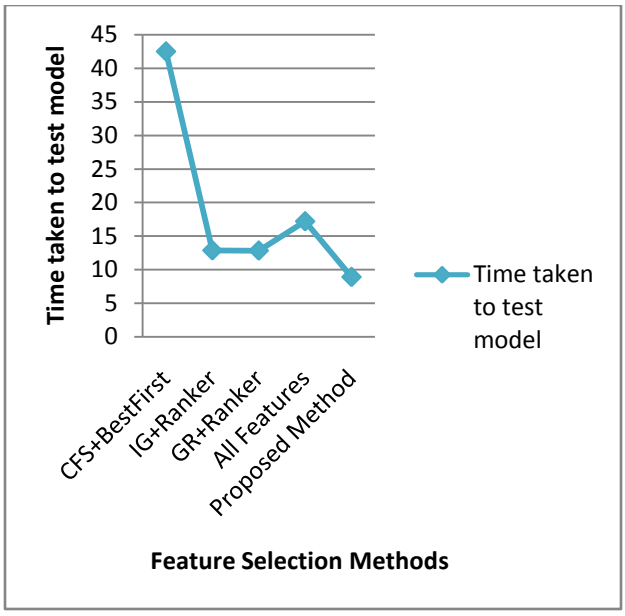

Figure 3: Time taken to test model

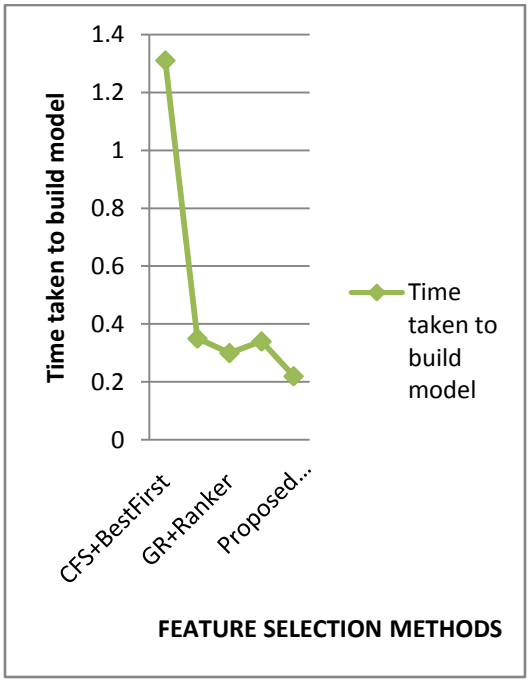

Figure 2: Time taken to build model

8. Conclusion and Future Work

We have proposed a new method for attribute selection by making use of the standard algorithms i.e. CFS, Information Gain, Gain Ratio and Genetic Algorithm. By using the proposed algorithm the result improves in terms of reduction in feature set, reduction in testing and training time and also gain increase in detection rate. Future work will include considering the 4 classes of attack .

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] Bezroukov, Nikolai, "Intrusion Detection (general issues)." Softpanorama: Open Source Software Educational Society. Nikolai Bezroukov, URL: http://www. softpanorama.org/Security/intrusion detection.shtml , 2003.

[4] Caruana,R. and Frietag,D. "Greedy Attribute Selection," Proc. 11th Int'l Conf. Machine Learning, pp. 28-36, 1994.

[5] Yeung, D.Y. & Ding, Y. (2003),"Host-based intrusion detection using dynamic and static behavioral models", Pattern Recognition, 36, 229-243.
[6] Mitra, P. et al. (2002),"Unsupervised Feature Selection Using Feature Similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 301–312.
[7] Blum, Avrim, L. and Langley, P.. "Selection of relevant features and examples in machine learning", Artificial Intelligence, 97(1-2):245–271, 1997.
[8] Kohavi, R. and John, G. (1997),"Wrappers for Feature Subset Selection. Artificial Intelligence", 97 (1-2), 273-324.
[9] Liu, H., Motoda , H. ," Feature Selection for Knowledge Discovery and Data Mining", Boston: Kluwer Academic, 1998.
[10] Kim, Y.,Street,W. and Menczer,F.(2000) "Feature Selection for Unsupervised Learning via Evolutionary Search," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 365-369.
[11] Das, S. (2001)," Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection", Proc. 18th Int'l Conf. Machine Learning, 74-81.
[12] Xing, E. et al. (2001)"Feature Selection for High-Dimensional Genomic Microarray Data", Proc.15th Int'l Conf.Machine Learning, 601-608.

[13] 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.
[14] H. GüneşKayacık, A. NurZincir-Heywood "Selecting Features for Intrusion Detection:
A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets".
[15] Shazzad, K. and Park, J. (2005)"Optimization of intrusion detection through fast hybrid feature selection" , IEEE.
[16] T.Pingjie, J.Rong-an (2010) "Feature selection and design of intrusion detection system based on k-means and triangle area support vector machine", IEEE.
[17] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning", Proc. 17th Int'l Conf Machine Learning, 2000, pp. 359-366.
[18] j.Han ,M Kamber, Data mining : Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers(2001).
[19] http://www.cs.waikato.ac.nz/~ml/weka/
[20]Lei Yu and Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learing Research 5(2004), pp1205-1224.

[21] M. Kudo, J. Sklansky, "Comparison of algorithms that select features for pattern classifiers", Pattern Recognition 33 (2000) 25-41.

[22] H. Pohlheim, "Genetic and Evolutionary Algorithms: Principles, Methods and Algorithms ", http://www.geatbx.com/doculindex.html.
[23] L.Y. Zhai, L.P. Khoo, and S.C. Fok, "Feature extraction using rough set theory and genetic algorithms and application for the simplification of product quality evaluation", Computers & Industrial Engineering, 2002, pp. 661-676.

M**eghaAggarwal**received herB.Tech degree with honors in Computer science and engineeringfrom UPTU university. She is pursuing M.Tech in computer science and engineering from Shardauniversity. 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.