INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ANALYSIS OF DIFFERENT FEATURE SELECTION METHODS IN INTRUSION DETECTION SYSTEM

S.Gnanasundari[1], P.Narendran[2],
[1] *Research Scholar, Department of Computer Science, Gobi Arts & Science College,*
gnanasundaribscit@gmail.com
[2]*Associate Professor & Head, Department of Computer Science, Gobi Arts & Science College,*
narendranp@gmail.com
Contact No: 8148784709

Abstract

In today's detection of era security threats that are usually said as intrusion, has become a really vital and significant issue in network, information and knowledge security. Extremely confidential information of varied organizations is gift over the network therefore so as to preserve that information from unauthorized users or attackers a powerful security framework is needed. Intrusion detection system plays a serious role in providing security to pc networks. AN Intrusion detection system collects and analyzes info from totally different areas at intervals a pc or a network to spot attainable security threats that embody threats from each outside furthermore as within the organization. The Intrusion detection system deals with great deal of information that contains numerous orthogonal and redundant options leading to inflated time interval and low detection rate. So feature choice plays a crucial role in intrusion detection. There is numerous feature choice strategies projected in literature by totally different authors. In this paper a comparative analysis of various feature choice strategies are evaluated in terms of detection rate, root mean sq. error and procedure time.

Key Words: Intrusion Detection, Comparative Analysis, KDDCup99 dataset, Feature Selection

1. Introduction

During the previous couple of years there's a dramatic increase in growth of pc networks. There are numerous personal additionally as government organizations that store valuable information over the network. This tremendous growth has expose difficult problems in network and knowledge security, and detection of security threats, ordinarily noted as intrusion, has become a really vital and significant issue in network, information and knowledge security. the safety attacks will cause severe disruption to information and networks. Therefore, Intrusion Detection System (IDS) becomes a crucial a part of each pc or network system. Intrusion detection (ID) may be a mechanism that has security for each computers and networks. The paper is organized into the subsequent sections. Intrusion Detection Systems is reviewed in Section a pair of. Section three offers the small print of the datasets employed in this comparative analysis. In Section four, totally different methodologies of feature choice in IDSs area unit mentioned. Connected analysis within the literature for feature choice ways is self-addressed in Section five. Section six conferred the results drawn from comparative analysis in tabular kind. Section seven concludes the discussion over comparative analysis.

2. Intrusion Detection System

An intrusion is an endeavour to compromise the integrity, confidentiality, availableness of a resource, or to bypass the protection mechanisms of a ADPS or network. James Anderson introduced the construct of intrusion detection in 1980 [1].It monitors pc or network traffic and establish malicious activities that alerts the system or network administrator against malicious attacks. Dorothy Denning projected many models for IDS in 1987 [2].

Approaches of IDS supported detection area unit anomaly based mostly} and misuse based intrusion detection. In anomaly based mostly intrusion detection approach [3], the system 1st learns the traditional behaviour or activity of the system or network to notice the intrusion. If the system deviates from its traditional behaviour then AN alarm is created. In misuse based mostly intrusion detection approach [4], IDS monitors packets within the network and compares with keep attack patterns called signatures. the most disadvantage is that there'll be distinction between the new threat discovered and signature being employed in IDS for detective work the threat. Approaches of IDS supported location of watching area unit Network based mostly intrusion detection system (NIDS) [5] and Host-based intrusion detection system (HIDS) [6]. NIDS detects intrusion by watching network traffic in terms of informatics packet. HIDS area unit put in regionally on host machines and detects intrusions by examining system calls, application logs, filing system modification and different host activities created by every user on a specific machine.

3. Datasets

The KDD CUP 1999 [7] benchmark datasets square measure utilized in order to judge completely different feature choice technique for Intrusion detection system. It consists of four, 940,000 association records. every association had a label of either traditional or the attack kind, with specifically one specific attack kind falls into one in every of the four attacks classes [8] as: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to native Attack (R2L) and inquiring Attack.

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

Probe Attack: These attacks mechanically scan a network of computers or a DNS server to search out valid information science addresses

.

Remote to Local (R2L) Attack: during this sort of attack associate degree assailant WHO doesn't have associate degree account on a victim machine gains native access to the machine and modifies the information.

User to Root (U2R) Attack: during this sort of attack an area user on a machine is ready to get privileges commonly reserved for the super (root) users. Each connection record consisted of 41 features and are labelled in order as 1,2,3,4,5,6,7,8,9,.....,41 and falls into the four categories are shown in Table 1:

Category one (1-9) :Basic options of individual transmission control protocol connections.

Category a pair of (10-22) : Content options inside a association urged by domain information.

Category three (23-31) : Traffic options computed employing a two-second time window.

Category four (32-41): Traffic options computed employing a two-second time window from destination to host.

Table 1:
Distribution of intrusion types in datasets

Dataset	Normal	Probe	DOS	U2R	R2L	Total
(-kdd data-10-perce	97280	4107	391458	52	1124	494020

4. Feature Selection

Due to the massive quantity of knowledge flowing over the network real time intrusion detection is sort of not possible. Feature choice will cut back the computation time and model complexness. Analysis on feature choice started in early 60s [9]. Primarily feature choice may be a technique of choosing a set of relevant/important options by removing most immaterial and redundant options [10] from the information for building economically good and efficient learning model [11].

Process of Feature Selection

Feature choice processes involve four basic steps during a typical feature choice technique [11] shown in Figure one. initial is generation procedure to get subsequent candidate set; other could be an analysis perform to judge the set and third one is a stopping criterion to come to a decision once to stop; and a validation procedure to envision whether or not the subset is valid.

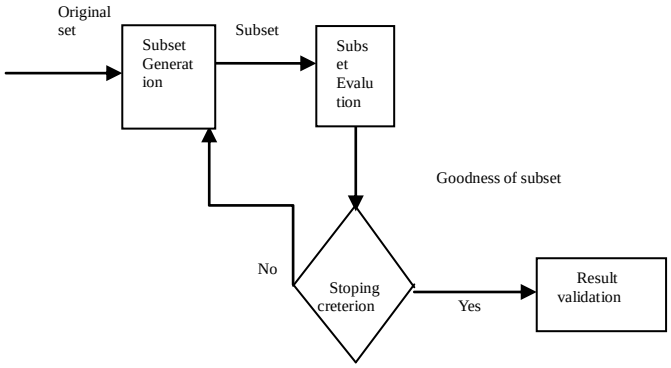

Fig1: Feature selection process [11]

Methods for Feature Selection:

Blum and Langley [12] divide the feature choice strategies into 3 classes named filter, wrapper and hybrid (embedded) technique.

Filter technique: Filter method [13] uses external learning formula to guage the performance of selected options.

Wrapper technique: The wrapper method [14 the training formula. It uses one preset classifier to evaluate options or feature subsets. Wrapper formula [15] uses a hunt formula to go looking through the house of doable options and value every set by running a model on the set. several feature subsets area unit evaluated supported classification performance and best one is chosen. This technique is a lot of computationally expensive than the filter technique [16] [14].

Hybrid technique: The hybrid method [16] [17] combines wrapper and filter approach to attain absolute best performance with a specific learning formula.

5. Background

In paper [18], a feature choice approach supported Genetic Quantum Particale Swarm optimisation (GQPSO) for network intrusion detection has been projected. Within the approach, choice and variation of genetic algorithmic program with QPSO algorithmic program area unit combined to make GQPSO algorithmic program. The projected methodology reduces redundant and inapplicable options. Experimental results show that classification detection rate and detection speed of GQPSO algorithmic program is on top of those of PSO and QPSO algorithms. Support Vector Machine (SVM) is employed as a classifier. In paper [19], a straightforward Genetic algorithmic program (GA) is employed to evolve weights for the options and k-nearest neighbour (KNN) classifier is employed as fitness operate of the GA and additionally as classifier. One advantage of the KNN methodology is that it's straightforward to use a weight to a feature of the info set. This weighted feature set has reduced noise gift within the information and improved levels of KNN classification. Top 5 graded options for every category area unit chosen. The result shown indicates a rise in intrusion detection accuracy. This paper [20] projected AN approach wherever genetic search strategies together with correlation square measure used for feature choice and system is employed as a classifier. a brand new AI paradigm referred to as the factitious system (AIS) was created supported human system. To implement a basic artificial system, four choices ought to be made: secret writing, Similarity live, choice and Mutation. Attributes square measure chosen supported correlation primarily based feature mistreatment genetic search. the chosen options square measure accustomed train the AIS algorithmic program and after tested. Within the paper 2 soft computing techniques for Network Intrusion Detection System (NIDS) square measure used. A genetic search approach was thought-about for correlation primarily based feature choice. Artificial system (AIS) primarily based classifier was accustomed classify the category labels over the chosen options. Results obtained show recall of ninety nine.7% for traditional information. Recall of three.5% was obtained for teardrop that had just one instance within the dataset. In paper [21], they projected a replacement hybrid feature choice methodology –a fusion of Correlation-based Feature choice, Support Vector Machine associated Genetic algorithmic rule –to confirm an best feature set. Correlation-based Feature choice (CFS) could be a filter methodology. It evaluates advantage of the feature set. A flow chart is given during this paper that describes the operating of the projected hybrid algorithmic rule. The hybrid feature choice methodology reduced the procedure resource whereas maintaining the detection and false positive rate among tolerable vary. The projected algorithmic rule additionally reduces the coaching time and testing time. Quicker coaching and testing helps to create light-weight intrusion detection system.

6. Study of feature selection methods

A number of feature choice algorithms square measure projected by varied authors. The aim of this work is to look at the assorted existing attribute choice ways in terms of detection rate and procedure time. Out of the overall forty one network traffic options, employed in detection intrusion, some options are going to be potential in detection intrusions. Thus the predominant options square measure extracted from the forty one options that square measure very effective in detection intrusions.

Attribute evaluators [22]:

Attribute authority is largely used for ranking all the options per some metric. Numerous attribute evaluators are on the market in wood hen. We have a tendency to used (Weka, 3.7.8) a learning machine during this work which has CfsSubsetEval, ChiSquaredAttributeEval, InfoGainAttributeEval and GainRatioAttributeEval.

a. CfsSubsetEval: Evaluates the value of a set of attributes by considering the individual ability of every feature in conjunction with the degree of redundancy between them. Subsets of options that are extremely correlative with the category whereas having low inter correlation with the opposite attributes are most popular.

b. ChiSquaredAttributeEval: Evaluates the value of Associate in Nursing attribute by computing the worth of the chi-squared datum with relation to the category.

c. GainRatioAttributeEval: Evaluates the value of Associate in Nursing attribute by mensuration the gain quantitative relation with relation to the category.

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

d. InfoGainAttributeEval: Evaluates the value of Associate in Nursing attribute by mensuration the knowledge gain with relation to the category.

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

Search Methods:

These strategies search the set of all attainable options so as to find the simplest set of options. Four search strategies which has BestFirst, GeneticSearch, GreedyStepwise and Ranker accessible in rail are employed in this work for comparison purpose.

a. Best First: These searches the area of attribute subsets by greedy hill ascension increased with a backtracking facility. Setting the amount of consecutive non-improving nodes allowed Controls the extent of backtracking done. Best 1st might begin with the empty set of attributes and search forward, or begin with the total set of attributes and search backward, or begin at any purpose and search in each directions (by considering all attainable single attribute additions and deletions at a given point).

b. Genetic Search: It performs enquiry exploitation the easy genetic algorithm.

c. GreedyStepwise: It performs a greedy forward or backward method search through the area of attribute subsets. Might begin with no/all attributes or from Associate in nursing discretionary purpose within the area. Stops once the addition/deletion of any remaining attributes ends up in a decrease in analysis. may also turn out a stratified list of attributes by traversing the area from one facet to the opposite and recording the order that attributes are elite.

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

7. Results and Discussions

We used we have a tendency weka (3.7.8) a learning machine to draw the comparative analysis. during this paper totally different combination of feature choice strategies square measure tried and that they embrace BestFirst + CfsSubsetEval, GeneticSearch + CfsSubsetEval, GreedyStepwise + CfsSubsetEval, Ranker + ChiSquaredAttributeEval, Ranker + InfoGainAttributeEval and Ranker + GainRatioAttributeEval. the small print of the mixtures and also the options hand-picked by every combination and their mental image is delineated in Table one, 2, 3, Figs. 2 and 3.

Table2:
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,

Table3:
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 bulid 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

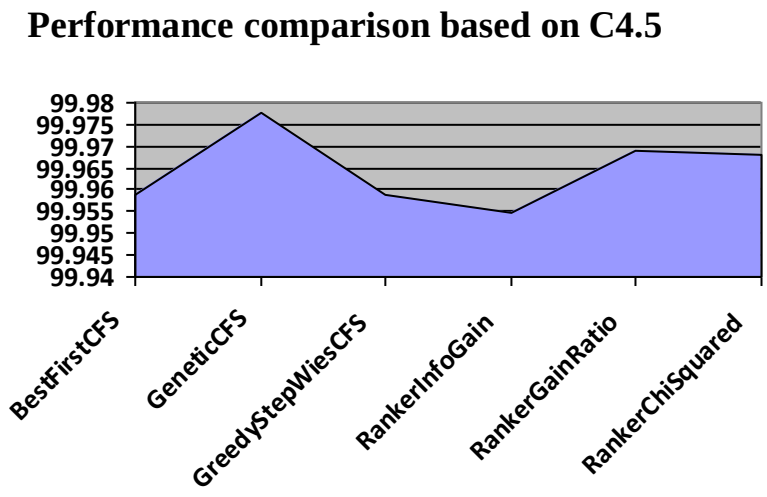

Fig2. Performance comparisons of various feature extraction algorithms based on C4.5

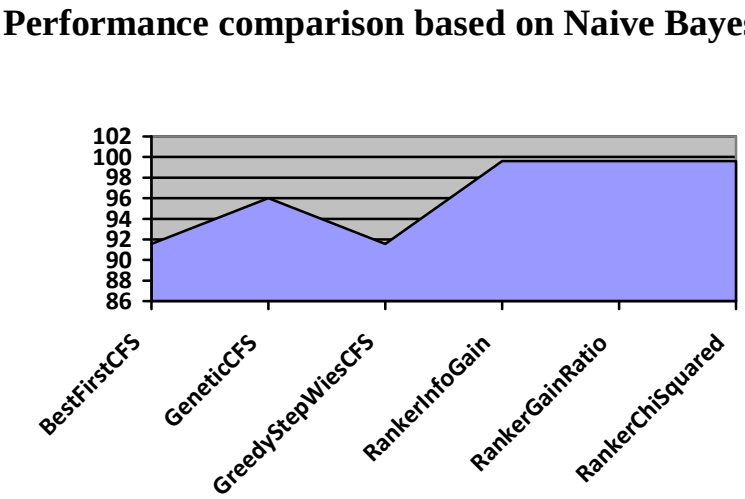

Fig3. Performance comparisons of various feature extraction algorithms based on Naive Bayes

8. Conclusion

In this paper a comparative analysis has been done on the premise of detection rate, process time and root mean sq. error. During this analysis six feature choice algorithms are used and their performance is evaluated victimization Naïve Bayes and C4.5(J48) classifier. options selected victimization Bestfirst+CFSSubsetEval and GreedyStepwise+CFSSubsetEval is same thus their performance is same. GeneticSearch+CFSSubsetEval performance is nice over different 2 and that we will say CFS performs best with genetic search.InfoGain, GainRatio and Chisquared are feature choice strategies that are supported ranking. thus on the premise of ranking we have a tendency to choose prime twenty five attributes from every of the 3 feature choice strategies so by doing analysis it's been determined that the performance of Ranker+GainRatioAttributeEval is nice in terms of detection rate however it takes additional testing and coaching time. Ranker+InfoGainAttributeEval takes less process time among all the feature choice strategies. during this paper 2 classifiers are used specifically Naive Bayes and C4.5 and it's been determined that Naive Bayes takes less time to check the dataset however longer in coaching the set whereas C4.5 will 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 Software Engineering‖, 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/attributeSelection.