ISSN:0975-9646

Reduces Unwanted Attribute in Intruder File Based on Feature Selection and Feature Reduction Using ID3 Algorithm

Uma Vishwakarma, Prof. Anurag Jain

Department of CSE,
RITS, Bhopal

Abstract-Reduction and selection of intruder attribute in intrusion detection system play an important role in process of detection. The huge number of attribute in intruder introduces a problem in detection process and increase more time in decision making process. Previous researchers used some standard techniques for feature reduction such as Principle component analysis (PCA),Pulse coupled neural network (PCNN) , but these methods are not consider some fixed no of feature at time of processing features from dataset. In this paper proposed a feature selection and feature reduction method based on improved Interactive Dichotomiser 3 (ID3) algorithm. The proposed algorithm select multiple feature for reduction and the reduce feature set participant the process of detection. For the process of performance evaluation used KDDCUP99 dataset and simulate in MATLAB software achieved 96.67% detection rate.

Keywords: - intrusion detection, feature selection and reduction, KDDCUP99 and ID3.

I INTRODUCTION

The performance of intrusion detection system depends on classification of unknown types of attacks. The detection of unknown types of attack is very difficult task due to large number of attribute and huge amount of network data for the improvement of unknown attacks feature reduction is important area of research. The reduction process reduces the large number of attribute and improved the detection of intrusion detection system. In the process of feature reduction various algorithm are used such algorithm are principle component analysis and neural network. The reduction process used PCA method this method is static reduction technique, reduces only fixed number of attribute. The fixed number of feature reduction process not justify the value of feature it directly reduces the feature. On the consideration of computational time feature reduction is also an important aspects, the reduces feature increase the processing of detection ratio. Many methods have been proposed in the last decades on the designs of IDSs based on feature reduction technique. For example silakari and saliendra[1] proposed a generic framework for intrusion detection based on feature reduction and ensemble based classifier. On the other hand genetic algorithm is directly applied for classification in the work of Jain and Upendra [2] applied information gain based feature reduction for intrusion detection. They used KDDCUP'99 dataset for comparing four machine learning algorithms and they found that J48 classifier outperforms over BayesNet, OneR and NB classifiers. Muda, Y. Yassin [3] also used KDDCUP'99 dataset for evaluating their K-Means and Naive Bayes based learning approach to carry out intrusion detection. Support Vector Machine (SVM) based IDS with Principal Component Analysis (PCA) dimension reduction is presented for intrusion detection in [4,5]. Z. Xue-qin et al. [6] proposed SVM IDS with Fisher score for feature selection. Zhang and M. Zulkernine [7] applied random forests for network intrusion detection. in this paper used ID3 algorithm for feature selection. ID3 is attribute based classification technique in decision tree. The selection of attribute in ID3 algorithm is entropy of information and gain of information. The increasing the sample selection area used radial biases function in ID3 algorithm. The sample selection process based on correlation coefficient of sampled data in attribute selection process. . In section II we describe feature of intruder file and ID3 algorithm. In section III proposed algorithm. In section IV discuss experimental result analysis. In section and finally conclude in section V.

II FEATURE SELECTION

The network traffic generate huge amount of traffic data in every few seconds, the processing of these data for firewall and intrusion detection system is very complex. The complex raw data is not formatted and standard relation for the process of filtration and classification. These original raw data process through KDD data mining tools and converted into connection. a connection justify the sequence of packet form source to distention. The process of conversion performs by Paxson algorithm. Finally get 41 features. These feature divided into four categories[3].

1. Basic Features: - These features are captured from packet headers only and without analyzing payload. Features 1 to 8 are in this category.
2. Content Features: - In this category original TCP packets analyzed with assistance of domain knowledge. An example of this category is number of "hot" indicators.
3. Time-based Traffic Features: - for capturing these types of features a window of 2 second interval is defined. In this interval, some properties of packets are measured. For example number of connections to the same service as the current connection in the past two seconds.
4. Host-based Traffic Features: - In this category instead of a time based window, a number of connections are

used for building the window. This category is designed so that attacks longer than 2 second can be detected.

The processing of feature and description of feature discuss in table 1,2 and 3 according to their description and data type

Table 1Basic features of individual TCP connections

Feature name	Description	Type
hot	number of ``hot'' indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of ``compromised'' conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
num_root	number of ``root'' accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an ftp session	continuous
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete

Table 2: Content features within a connection suggested by domain knowledge.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN'' errors	Continuous
rerror_rate	% of connections that have ``REJ'' errors	Continuous
same_srv_rate	% of connections to the same service	Continuous
diff_srv_rate	% of connections to different services	Continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	Continuous
srv_serror_rate	% of connections that have ``SYN'' errors	Continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	Continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Table 3 Traffic features computed using a two-second time window.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN'' errors	continuous
rerror_rate	% of connections that have ``REJ'' errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
srv_serror_rate	% of connections that have ``SYN'' errors	continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Feature selection is an important data processing step earlier to applying a detection algorithm. It is a process of determining whether a feature is relevant or not for a particular algorithm. Using effective features to propose algorithm not only can reduce the data size but also can improve the performance of the detection and enhanced the performance of intrusion detection system. One of the major problems in feature reduction is to select effective attributes that have the best discrimination ability between the groups. There are two common approaches for feature reduction: Wrapper and Filter[13,14]. A Wrapper method selects feature subset based on the performance of the learning algorithm that is going to be used. Wrapper method is totally dependent on the learning algorithm. On the other hand Filter methods evaluate features according to statistical characteristics of the data only without the involvement of any learning algorithm. The wrapper approach is generally considered to produce better feature subsets but runs much more slowly and requires more computing resource than a filter method.

ID3[12] is an attribute based classification technique in data mining. The flexibility of ID3 algorithm in case of small data is very high. The process of ID3 Algorithm based on information entropy of attribute. in the process of feature selection ID3 are used as find common feature for selection process. The ID3 algorithm is the basic algorithm of decision tree induction, it generates decision tree by means of defeating in detail from the top to the bottom.

Algorithm:

The algorithm is as follows:

ID3 (data, Target_Attribute, Attributes)

- Create a root node for the tree
- If all data are positive, Return the single-node tree Root, with label = +.
- If all data are negative, Return the single-node tree Root, with label = -.

- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the data.
- Otherwise Begin
  - A = The Attribute that best classifies data.
  - Decision Tree attribute for Root = A.
  - For each possible value, v_i, of A,
    - Add a new tree branch below Root, corresponding to the test A = v_i.
    - Let Data(v_i) be the subset of data that have the value v_i for A
    - If Data(v_i) is empty
      - Then below this new branch add a leaf node with label = most common target value in the data
    - Else below this new branch add the subtree ID3 (Data(v_i), Target_Attribute, Attributes – {A})
- End
- Return Root

III Proposed Method

In this section described a proposed method for improved ID3 algorithm for feature reduction come classification technique. The huge amount of feature process through our sample selection process, the sample selection process used correlation factor for estimated feature value for reduction process. The radial biases function (RBF) is gusssain nature. The nature of mixture data correlation of attribute used in ID3 algorithm. The combination of RBF and ID3 algorithm perform well feature reduction cum classification process over intrusion data. The RBF function incases the size of sample selection. The incasing size of sample selection incases the range of feature attribute of intruder data. RBF function is creating for sample selection for reduces and unreduced categories data sample for dealing out of ID3 classification. The input processing of training phase is data sampling technique for classifier. Single-layer RBF networks can potentially learn virtually any input output relationship; RBF networks with single layers might learn complex relationships more quickly. The function of ID3 creates forward networks. The network-layer network also has connections from the input to all cascaded layers. The additional connections might improve the speed at which the network learns the desired relationship. RBF artificial intelligence model is similar to feed-forward back-propagation neural network in using the back-propagation algorithm for weights updating, but the main indication of this network is that each layer of neurons related to all previous layer of neurons. The process of feature reduction and classification steps given below

1. input the dataset
2. estimate the feature correlation attribute as

$$Rel(a,b) = \frac{cov(a,b)}{\sqrt{var(a)\times var(b)}}$$ Here a and b the feature attribute of input data

3. the estimated correlation coefficient data passes through RBF function as

$$x(t) = w0 + \sum_{j=1}^{total\ data} wj\ exp\left(\frac{-(total - xj)}{\sigma^2}\right)$$

4. create the relative feature difference value
   $Rc = \sum_{k=1}^{r}\sum_{l=1}^{m}(hl - h)(elk - et)$
5. After sampling of feature data get reduces set of feature attribute of feature matrix.
6. generate feature attribute of each matrix
7. compute the entropy of feature attribute for the root node
   $Entropy(D) = -\sum_{i=1}^{N} pi \log pi \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$
8. compute the information gain of feature attribute
   $FB(v) = \sum_{j=1}^{N} Pj[pi \log pi] \ldots\ldots\ldots\ldots (2)$
9. compute the gain of each feature attribute
   Gain (v) =Entropy (D)-FB (v)
10. determine maximum gain of feature value and split encode feature in Gaussian form
11. part ion the root node and leaf node
12. data are classified
13. estimate the classification ratio
14. exit

IV Experimental Result Analysis

Evaluate the performance of proposed algorithm intrusion detection; using KDD'99 attack datasets [18]. In KDD99 dataset these four attack classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that tabulated in Table I. The 1999 KDD datasets are divided into two parts: the training dataset and the testing dataset. The testing dataset contains not only known attacks from the training data but also unknown attacks. Since 1999, KDD'99 has been the most wildly used data set for the evaluation of anomaly detection methods.DARPA'98 is about 4 gigabytes of compressed raw (binary) tcp-dump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. For each TCP/IP connection, 41 various quantitative (continuous data type) and qualitative (discrete data type) features were extracted among the 41 features, 34 features (numeric) and 7 features.

Table4. Different types of attacks in kdd99 dataset

4 Main Attack Classes	**22 Attack Classes**
Denial of Service (DoS)	back, land, neptune, pod, smurt, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf,spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Probing(Information Gathering)	ipsweep, nmap, portsweep, satan

To analysis the different results using some standard parameter such as Precision- Precision measures the proportion of predicted positives/negatives which are actually positive/negative. Recall -It is the proportion of actual positives/negatives which are predicted positive/negative. Accuracy-It is the proportion of the total number of prediction that were correct or it is the percentage of correctly classified instances. False-negative rate (FN) is the percentage that attacks are misclassified from total number of attack records. False-positive (FP) is the percentage that normal data records are classified as attacks from total number of normal data records. Below we are showing how to calculate these parameters by the suitable formulas. And also, below we are showing the graph for that particular data set [17].

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

Accuracy = $\frac{TP+TN}{TP+TN+FN+FP}$

FPR= $\frac{FP}{FP+TN}$, FNR= $\frac{FN}{FN+TP}$

The assessment metrics are computed for testing dataset in the testing phase and the obtained result for all attacks and normal data are given in table 2, which is the overall classification performance of the proposed system on KDD cup 99 dataset. For the estimated result discuss in the form of number of reduces attribute and classification rate and execution time for classification process.

Table 5. Result table

No.of Attribute	Attribute name	Classification ratio	Classification time(ns)
41	All feature	98.957	**1.8983**
9	Port_type Service Flag Sa_srv_rate Diff_srv_rate Srv_di_ho_ra Dst_host_cou, Dst_host_srv_count Dst_ht_sa_srv_rate.	96.671	**1.348**
7	Port_type Service Sa_srv_rate Dst_host_cou, Dst_host_srv_count Dst_ht_sa_srv_rate. Srv_di_ho_ra	96.671	**1.274**
5	Port_type Service Sa_srv_rate Dst_host_cou, Dst_ht_sa_srv_rate.	96.671	**1.273**

Figure 1 shows that discretion of number of attribute and classification rate the number of feature attribute decease the classification rate increases.

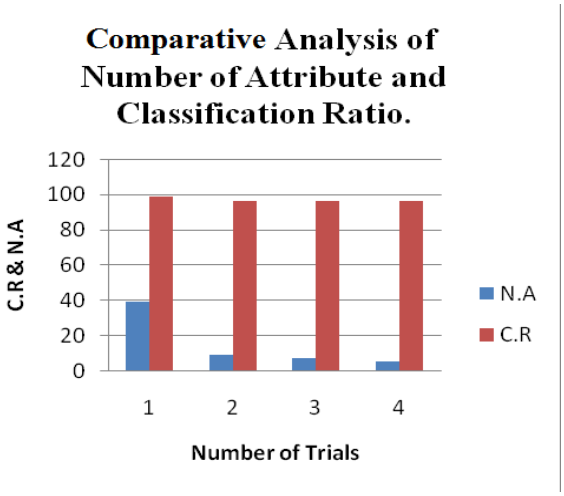


Figure 2 shows that discretion of number of attribute and classification rate the number of feature attribute decease the classification rate increase.

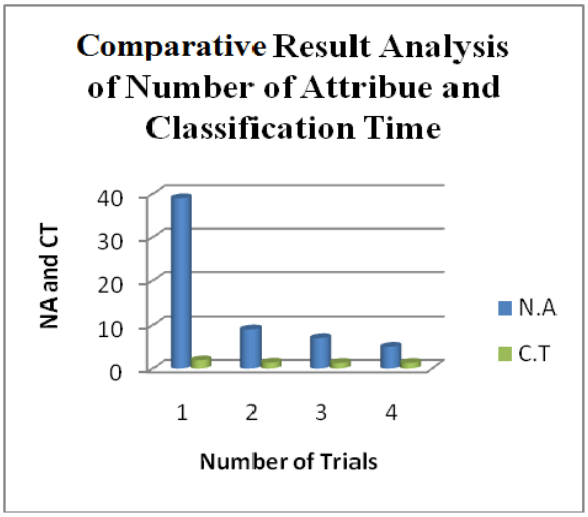


Figure 3 shows that the comparative valuation of three factor such as number of attribute .classification ratio and classification time.

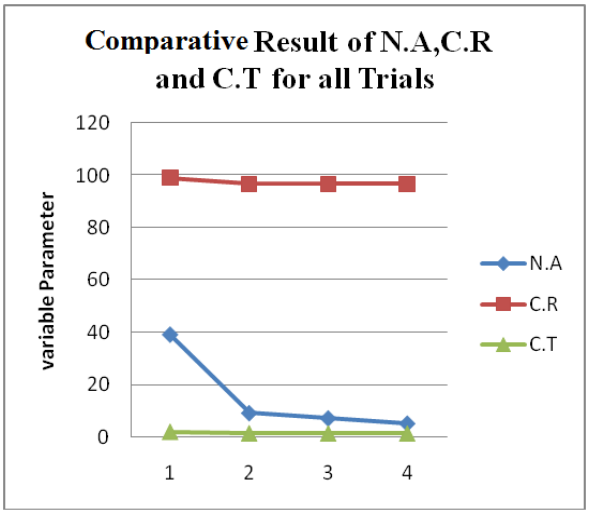
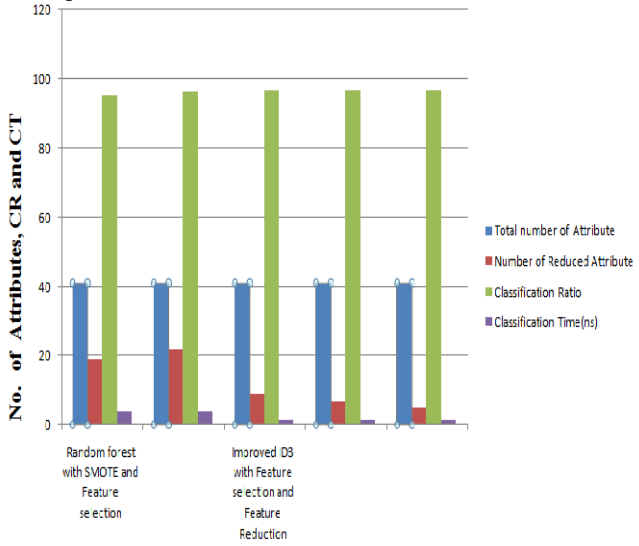


Table 6 shows that the comparative variance of time over number of attribute between SMOTE with Random Forest and Proposed Method Feature Selection and Feature Reduction With ID3 if number of attribute decrease the execution time also decrease and the performance of IDS are improved.

.

Algorithm Applied	Total number of Attribute	Number of Reduced Attribute	Classification Ratio	Classification Time ns
Random forest with SMOTE and Feature selection	41	19	95.30	3.913
		22	96.30	3.943
Improved ID3 with Feature selection and Feature Reduction	41	9	96.671	1.348
		7	96.671	1.274
		5	96.671	1.273

Figure 4 shows that the comparative variance of time over number of attribute between SMOTE with Random Forest and Proposed Method Feature Selection and Feature Reduction With ID3 if number of attribute decrease the execution time also decrease and the performance of IDS are improved.



V CONCLUSION AND FUTURE WORK

In this paper proposed a feature based intrusion data classification technique. The reduces feature improved the classification of intrusion data. The reduction process of feature attribute performs by RBF function along with feature correlation factor. The proposed method work as feature reducers and classification technique, from the reduction of feature attribute also decrease the execution time of classification. The decrease time increase the performance of intrusion detection system. Our experimental process gets some standard attribute set of intrusion file such asport_type, service, sa_srv_rate, dst_host_count,dst_host_sa_srv_rateThese feature attribute are most important attribute in domain of traffic area. The classification rate in these attribute achieved 98% In future improved the classification rate approx 100% used these attribute using another sampling technique

REFERENCES:-

[1] Shailendra Singh, Sanjay Silakari "An Ensemble Approach for Cyber Attack Detection System: A Generic Framework" 14th ACIS, IEEE 2013.

[2]Jain and Upendra "An Efficient intrusion detection based on Decision Tree Classifier using feature Reduction", International Journal of scientific and research Publications , Vol. 2, Jan. 2012.

[3]Muda, Y. Yassin, M.N. Sulaiman and N.I. Udzir, " A K-Means and Naive Bayes Learning Aproach for Better Information Detection", Information Technology journal, Asian Network For scientific Information publisher, Vol. 10 , 2011.

[4]Kausar , B.B Samir1, S.B Sulaiman, I. Ahmad , and M. Hussain, "An Approach towards Intrusion Detection using PCA Feature Subsets and SVM", Proc. International Conference on Computer & Information Science, IEEE Press, Jun. 2012.

[5]Koc, T. A. Mazzuchi, and S.Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Expert Systems with Applications: An International Journal, Vol. 39, Dec. 2012.

[6]Z. Xue-qin, G. Chun-hua and L. Jia-jin, "Intrusion detection system based on feature selection and support vector machine", Proc. First International Conference on Communications and Networking in China (ChinaCom '06), Oct. 2006.

[7] Zhang and M. Zulkernine, "Network Intrusion Detection using Random Forests", School of Computing Queen's University, Kingston Ontario, 2006.

[8]Shafigh Parsazad, Ehsan Saboori, Amin Allahyar" Fast Feature Reduction in Intrusion Detection Datasets" in MIPRO, May 21-25, 2012

[9]Abebe Tesfahun, D. Lalitha Bhaskari" Intrusion Detection using Random Forests Classifier With SMOTE and Feature Reduction" in International Conference on Cloud & Ubiquitous Computing & Emerging Technologies,2013

[10]Hachmi Fatma and Limam Mohamed "A two-stage technique to improve intrusion detection systems based on data mining algorithms" in IEEE,2013

[11]Li, "Using Genetic Algorithm for Network Intrusion Detection", Proc. the United States Department of Energy Cyber Security Group 2004 Training Conference, May 2004.

[12]Liu Yuxun and Xie Niuniu "Improved ID3 Algorithm" in IEEE, 2010

[13]Aslıhan Ozkaya and Bekir Karlık "Protocol Type Based Intrusion Detection Using RBF Neural Network" in International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (3) : Issue (4) : 2012.

[14]V.Venkatachalam and S.Selvan "Intrusion Detection using an Improved Competitive Learning Lamstar Neural Network" in IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.2, February 2007

[15] John Zhong Lei and Ali Ghorbani "Network Intrusion Detection Using an Improved Competitive Learning Neural Network" in Proceedings of the Second Annual Conference on Communication Networks and Services Research IEEE.