International Journal of Scientific Engineering and Technology
Volume No. 6, Issue No. 9, PP : 324-326
ISSN : 2277-1581
1 Sept. 2017

Systematic Feature Selection Based on Information Gain in Intrusion Detection Systems

Calpephore NKIKABAHIZI , Dr Wilson Cheruiyot , Dr Ann Kibe
Department of Information Technology, Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Corresponding Email : calponkika@yahoo.fr

***Abstract**- Network traffic has increasing marginally due to the availability of internet used, and cause the overload of dataset, and making data not be understandable. The monitoring of activities on internet using Intrusion Detection Systems (IDS) has been one of essential network infrastructure to ensure the security of internet. This IDS has been implemented based on internet features, therefore some of these features are irrelevant and the correspondents instances are redundant and inconsistent. . Feature selection is one of the most important preprocessing stages in data mining and knowledge engineering to overcome the problem of many variables, instances redundancy and inconsistency which make the problem not being approachable. This paper discusses systematic feature selection based on Information Gain to find the relevant subset of features which has effect on targets.*

Key words: Feature selection, Feature reduction, neural networks, intrusion detection systems

1. Introduction

Data mining for intrusions and prevention in networking has a great a potential to discover patterns of program , user activity, and determine what set of events indicate an attack (Ashok et al. , 2011).

To improve the quality of the pattern mined and/or the time for the actual mining, the different researchers applied data preprocessing techniques, such as data cleaning, data integration, and dimensionality reduction based on feature reduction and feature selection. This improvement which based on the philosophy of Coase (1991) , "if you torture the data for long enough, in the end they will confess'', is the way of searching for solution to emphasizing network security through reducing false alarm and time cost in IDS during monitoring malicious activities on network..

The feature selection dates back to the 1960's and its goal is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. This process is done without much more compromising information content. The mining on the selected set of attributes has additional benefits on dataset such as making the patterns easier to understand while interpreting, enhancing the classification accuracy (Huan & Lei , 2005), and learning runtime (Han & Kamber, 2001).

This paper presents use of a filter namely information gain for feature selection whereby the relevant features have been selected systematically are provided as input to neural networks which is a supervised classifier .

1.1 Information gain

According to Guyon and Elisseeff (2003), information gain (IG) known as mutual information determines which attribute in a given set of training features vector is most useful for discriminating between the classes to be learned and tries to find a subset of the original variable. It is one of three selection strategies: filter, wrapper and embedded approaches. The selection techniques are employed to select relevant and information features or to select features that are useful to build a good predictor. Information gain is based on Shannon's mathematical theory and communication , and depends on entropy, which is a measure of unpredictability of information, and rank the features that affect the data classification and p_i is the probability of i^{th} class in the given set of attributes (Gray, 2013).

I.G= entropy (Parent)-[average entropy (child)]

where $entropy = \sum_i -p_i \log_2 p_i$ (1.1) and

$$p_i = \frac{\#Class_i}{entity\ population} \quad (1.2)$$

According to Maher and Ulrich (2012), IG handles only discrete values ,therefore it is essential to transfer the continuous

DOI : 10.5958/2277-1581.2017.00049.3

values into discrete values. Given the two random variables X and Y, $I\left(X,\, Y\right)$ is the information gain of X with respect to the class attribute Y. When Y and are discrete variable that takes values in $\left\{y_i,\ldots\ldots.y_t\right\}$ and $\left\{x_i,\ldots\ldots.x_t\right\}$ with probability distribution function $P\left(x\right)$; then the entropy of X is given by $H(x) = -\sum_{i=1}^{t} P(X = x_i)\log_2(P(X = x_i))$ (1.3) Or the average information is expected value of I(x) over instance of X, $H(X) = E_X(I(x))$ (1.4), information I from the message X.

Hence the I.G for feature F on the dataset D in

$$IG(D,F) = H(D) - \sum_{attr=value}\left[\frac{|D_{attr}|}{|D|} * H(D_{attr})\right] \quad (1.5),$$

where value (F) is the set of all possible values of F, D_{attr} is the subset of D has a value attr. H(D)=entropy of class attribute and (.) donates cardinality (Schraudolph, 1995).

1.2 Basic concepts neural network

The term neural network refers to a network of biological neurons, or to artificial neural networks which composed by artificial neurons.

Artificial neural networks, also ,known as connectionist or parallel distributed processing system , are machine learning models implemented using a computation frameworks developed primarily to understand and simulate neural networks (Rumrlhart,1986).

Architecture of artificial neural networks

Neural network has input layers (visible) that interacts with the environment, hidden layers which do not interacts with the environment and output layers. Each layer in a neural network is composed of several nodes; each has associated activation status that is a function of the nodes input value. Artificial neural network has the connections called weight and are represented by the numbers and by directed edges, and get by training methods (Sharma & Prabin, 2011).

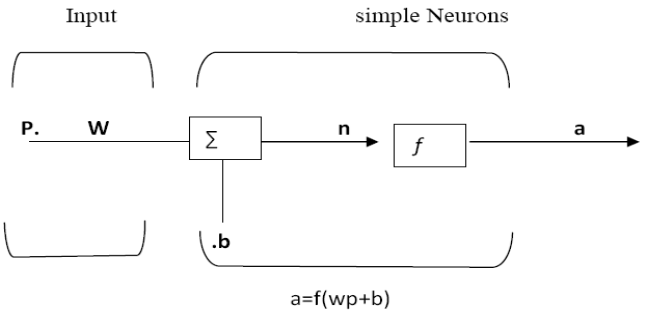
Based on the connectivity pattern, Artificial neural network can be grouped into two categories: Feed –forward networks where graphs has no loop or it allows signals to travel in one way, from input to output , and Feedback (recurrent) networks which signal travelling in both direction by introducing loops in networks. The various algorithm of neural network training are Hebb, Delta, Kohonen, and backpropagation computation. The mostly used BP computation is the error backpropagation algorithm proposed by Rumelhart in 1985 (Sharma & Prabin, 2011). Artificial neural network has a learning logarithms which learns rules used to adjust the weight. The used rule is of Habbian. The Hebbian rule (hebb, 1949) for learning in simple neural models dictates that “if two connected neurons (or nodes) are simultaneously in an active state, the connection between them should be strengthened”

Neural networks learning

Learning in NN is performed by iteratively modifying weights (w_i) such that the desired output is eventually produced by the networks, with the minimum amount of error. Initial small weights are updated gradually. And w_i is adjusted weight of i^{th} neuron .

Neural networks can classify pattern quickly once they know the value of weight, by performing simple operation such as multiplication and addition $y = f\left(\sum_{i=1}^{n} w_i x_i + bi\right)$ (1.6)

Figure 1.1 Learning process with simple neurons



In learning, you give, Initial weights, w_i initial output and initial bias “Ѳ”. Using this technology the system itself will adjust w_i, in the given activation function, such us linear threshold, parabolic or logistics to verify the approximate solution. Whenever the networks output is not closed to the desired output, a change is occurs in the direction that minimizes the error. Learning in network means finding a set of weights that minimizes the overall of error (Zhang, Fengli & Dan, 2013).

Backpropagation of error

Backpropagation is an abbreviation of backward propagation of errors. The backpropagation algorithm consists of the propagation of errors beginning at the output layer, through hidden layer, and so on to input layer, in backward direction.

The update of weight is done at each layer, and the change of weight is proportional to the derivative of errors with respect to the incoming weights. With the method, the system will adjust quickly the changed to find the desired solution. In learning by propagation, the weight w_{ij} connecting to units in the output layer are modified to the standard delta rule. Backpropagation is understanding how changing the weight and biases in network changes. Ultimately this means computing the partial derivative. In order to perform gradient descent each weights change between units I and J, and then Δw_{ij}

$$\Delta w_{ij} = \frac{\partial E}{\partial w} = \eta \delta_{ij} o_{pi} \quad , \delta_{ij} = (t_{pj} - o_{pj})(o_{pj})(1 - o_{pj}) \quad (1.7)$$

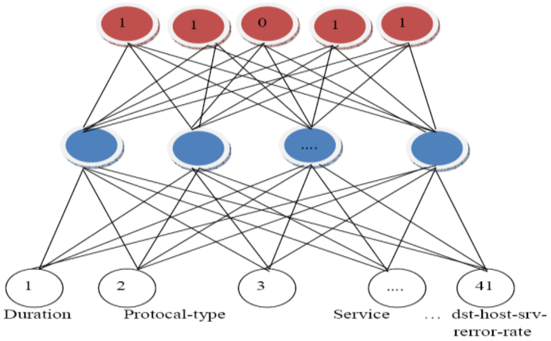
Where
t_{pj} : *the desired output*
o_{pj} : *the output calculated by the networks*

η : *the learning* rate
E : *the Error*

The problem raise in the Multi layer neural networks is to find an appropriate algorithm to estimate the weights in nonlinear function .The target of output layer is well defined, so that the use of delta rule can be used to this layer, but not at hidden layer. The update in intermediate layer using delta rule is an obstacle. This problem is solved by propagation algorithm, a generalization of the delta rule for multilayered neural networks.

According to Bhavin (2013) , the advantages of BPNN are to support the high speed in classification, to be used for linear as well as non linear classification and to support multi class classification. BPNN has disadvantages such as training time for BPNN that is high, suffering from local minima and the Structure of BPNN which is high complex, and training time

Figure 1.2: Multilayered neural networks for diagnosis any attack



a. No attack =0 (Normal)
b. Attack =1 (DOS,U2R,R2L,Probe)

In multilayer neural network, with n layers, we have composite of activation functions.

$$a^1 = f^1(Iw_{11}p + b_1), \quad a^2 = f^2(lw_{21}a^1 + b_2),$$
$$a^3 = f^3(lw_{32}a^2 + b_2)...... \quad a^n = f^n(lw_{n,n-1}a^{n-1} + b_n)$$

$$a^n = f^n(lw_{n(n-1)} f^{n-1}(lw_{n-1,n-2} f^{n-2}(......(lw_{21} f^1(iw_{11}p + b_1) + b_2) + b_3) + ... + b_{n-1}) + b_n) \ (1.8)$$

Where a^n is the output at n^{th} layer,f^n is the activation function, b_n is the bias ,$lw_{n,n-1}$,iw_{11} are weights connecting neurons from layer n-1 to layer n and from input to layer one respectively.

2. Related work

Wang and Goton (2009) conducted a research on feature selection using rough set, and deal with the problem raised while using microarray data analysis for cancer classification. Here the researcher selected informative genes from thousands or tens thousands to a small number of genes and therefore the results have performed well using simple rule.

Yasmen , Ehab and Ghada (2015) proposed a hybrid feature selection algorithm based on CFS and Information gain to reduce the number of features. Their research conducted in NSL-KDD dataset, and then the reduced dataset was trained by a naïve Bayes classifier using the adaptive boosting technique (AdaBoost.M1) which is showed to greatly enhance the classifier performance as well as decrease the false positive rate.

Sudhakar and Manimekalai (2015) experimented two filtering techniques the information gain and Principle Component Analysis filtering for feature selection in the given dataset to predict the heart disease symptoms, thereafter the results revealed that the techniques improves accuracies than others.

Zahra , Harounabadi and Mansour (2013) conducted a research on feature selection using information gain and Symmetric Uncertainty to select the relevant features for high performance of IDS using naïve Bayes classifier . The outputs showed that the proposed techniques performed more than others.

3. Methodology and Material

3.1 Pre-processing

Table 3.1 Attributes in IDS

Average Rank	Attribute	Average Rank	Attribute
1	Duration	22	is guest login
2	Protocal-type	23	count
3	Service	24	srv-count
4	Flag	25	serror -rate
5	src-bytes	26	srv-serror-rate
6	dst-bytes	27	rerror-rate
7	land	28	srv-rerror rate
8	wrong-fragment	29	same-srv rate
9	urgent	30	diff-srv-rate
10	hot	31	srv-diff-host rate
11	num- failed –logins	32	dst-host count
12	logged in	33	dst-host-srv-count
13	lnum-Compromise	34	dst-host-same-srv-rate
14	lroot-shell	35	dst-host-diff-rate
15	lsu-attempted	36	dst-host-same-src-port-rate
16	lnum-root	37	dst-host-srv-diff-host-rate
17	lnum-file creation	38	dst-host-serror-rate
18	lnum-shells	39	dst-host-srv-serror-rate
19	lnum-access files	40	dst-host-rerror-rate
20	lnum-outbound-cmds	41	dst-host-srv-rerror-rate
21	is host login		

IDS have 41 (full) attributes and 57297 instances in different forms, continuous, discrete, and /or symbolic. In this paper, a compression methods has been achieved through the normalization which started by encoding of text attributes to numerical value, $Protocal-type=\{1,2,3\}$, $Service\ value=\{1,2,3,...,65\}$, $Flag\ value=\{1,2,3,...,11\}$,and $Class\ attribute=\{1,2,3,4,5\}$. The encoding has been followed by Scaling values to (0 ,1) using firstly Minimum maximum normalization $n_v = f_2(v) = \dfrac{v-\min(v)}{\max(v)-\min(v)}$,secondly by statistical normalization or Zero mean normalization , $n_v = f_3(v) = \dfrac{v-\mu}{\sigma}$, and the lastly by decimal normalization $n_v = f_1(v) = \dfrac{v}{10^e}$, where = log(max(v)). The next phase, researcher used truncation function for Lossless size reduction whereby the dataset has been passed to size reduction unit, which replaces n.00 with n, where n is integer. This replacement reduces the size of the dataset; and therefore the false alarm rate was not been affected. By comparing the three techniques of normalization, the results showed that a MinMax method is a winner than other methods in terms of small size, accuracy and more speed with respect to time spent in building a model.

3.2 Feature selection

IDS dataset has 41 full attributes, whereby some of them may be irrelevant or redundant. The researcher picked out the useful features to improve the classifier accuracy with less time consuming. This work has been done on winner selected normalized dataset.

In this research, the information gain is used as technique to reduce the number of features. The feature which does not have much effect on the data classification has very small information gain and can be ignored without affecting the detection accuracy of a classifier. To assess its effect, the systematic selection based on average merit of IG, and grouped into range according to approximate values. The neural network has been used as classifier, and the number of hidden neuron has been found by consisting on optimal number of samples by Siddhartha (2008), whereby $samples = 100 * weights$, and the number of weights, $W = H*(I+O)+H+O$ where H represents number of neurons in hidden layer, I and O stand respectively for Input and Output. Therefore the number neurons in hidden layer is given by $H = \frac{\frac{sample}{100} - O}{I + O + 1}$, and learning rate has been fixed at 0.3

4. Result and discussion

The experimental results shows a ranked list of selected features using IG : 3, 37, 12, 35, 33, 34, 27, 28, 29, 40, 41, 2, 31, 32, 25, 30, 22, 26, 38, 10, 1, 13, 16, 8, 39, 17, 6, 5, 11, 14, 19, 36, 18, 4, 15, 9, 7, 23, 24, 20, 21 : 41

Table 4.1: Performance and time cost of system by different number of features

Average merit	Number of features	Perfor-mance(%)	RMSE	Time (Seconds)
All	41	99.45	0.0438	299.4
>0%	39	99.51	0.0415	300.94
$>10^{-3}$	**34**	**99.5**	**0.0417**	**280.74**
$>10^{-2}$	23	99.44	0.0457	279.76
$>2*10^{-2}$	20	99.39	0.0469	317.87
$>3*10^{-2}$	17	99.23	0.0523	319.8

The above table shows that the dataset with 34 features is the best one due to two the following criteria, one it performs more than the dataset with full attributes ,and second one it processes very fast in building model comparing to it. Others do not fit the two conditions in the same time. Remember that, if the classification performance varies in decreasing as the features are reduced, therefore those features have effect on class label.

Conclusion

This paper has discussed the systematic way of picking the variables which have effect on class label using IG, and the subset of 34 attributes shows a great effect .The results have been found using the normalized dataset and its performances by the neural network classifier has been found by different number of IDS inputs. The researcher recommends the future work to use the selected features on hybrid of classifiers.

References

i. Ashok, C.; Naresh, D. H. & Rohini, B. (2011). Data Mining Techniques for Intrusion Detection and Prevention System. International Journal of Computer Science and Network Security 11(8), 200-203.

ii. Bhavin , S. & Bhushan, H.T.(2012). Artificial neural networks based Intrusion Detection System: Survey International Journal of Computer Application 6, 13-18

iii. Coase,R.,H. (1991).Nobel Prize in Economics in 1991

iv. Han,J. & Kamber, M. (2001).Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffm ANNPublishers

v. Huan L., Lei Y. (2005) . Toward Integrating Feature Selection Algorithms for Classification and Clustering , IEEE Transactions On Knowledge and Data Engineering, 17,(4)

vi. Maher, S. & Ulrich, B. (2012).Mining techniques in network security to enhance intrusion detection system . International journal of network security & its application 4(5),51-66.

vii. Rumrlhart, (1986). Backpropagations : Theory, architectures, and application. Lawrence Eribaum Associates, New Jersey UK.

viii. viii. Schraudolph, N.N. (1995). Optimization of entropy with neural networks. Doctor of philosophy in cognitive science and computer science. University of California, san Diego.

ix. Sidhartha,B. (2005).University of Illinois at chichago,-UIC, class note of IDS. Data mining for business .

x. Sudhakar, K. & Manimekalai, M. (2015). Propose a Enhanced Framework for Prediction of Heart Disease. Journal of Engineering Research and Applications 5, (4), 1-6

xi. Wang, X. & Gotoh, O. (2009). Accurate molecular classification of cancer using simple rules. BMC Medical Genomics, 64(2),1-23.

xii. Yasmen W., Ehab ,E. & Ghada ,E. (2015). Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction. International Journal of Computer Science 12, (3), 255-262

xiii. Zahra, K.,Harounabadi, A., Mansour, M. (2013). Feature Ranking in intrusion Detection Dataset using Combination of Filtering Methods, International Journal of Computer 78 (4) , 21-27.

xiv. Zhang ,Fengli, & Wand, D. (2013). An effective feature selection approach for network intrusion detection system, networking, architecture and storage (NAS) . IEEE eight international conference on IEEE.