International Journal of Engineering and Manufacturing Science.
ISSN 2249-3115 Volume 8, Number 1 (2018) pp. 167-175

http://www.ripublication.com

Analysis of Feature selection algorithms for Naïve Bayes classifier using NSL-KDD

Vinutha H.P.[1], Dr.Poornima B[2]

Assistant Professor, Dept. of CS&E, BIET, Davangere, Karnataka, India[1]
Professor& Head, Dept. of IS&E, BIET, Davangere, Karnataka, India[2]
vinuprasad.hp@gmail.com[1], poornimateju@gmail.com[2]

Abstract

Growth of internet in our daily life and dependences on the network, the network security is becoming a major issue. In order to secure our network, it is necessary detect the attacks over the network. In this paper we tried to analyze the NSL-KDD dataset using different feature selection algorithms. By applying different filter methods on this NSL-KDD dataset we tried to remove the unused and least used attributes. Experimental result is shown by performing Naive Bayes classifier in WEKA environment.

Key words: Filtering method, NSL-KDD dataset, Naïve Bayes, WEKA.

1. INTRODUCTION

Day by day networks are becoming more complex. The growing need of internet made us to depend on computer network so the network security plays a major role. Therefore, we need to find a best possible way to protect our systems. The security of the system is compromised when the intrusion takes place. Intrusion Detection System (IDS) is the system that automates the processes of monitoring the events occurring in the network for the sign of security of the network. Data mining based IDS are used for this purpose. A data mining based IDS uses machine learning and data mining algorithms on a data set. Data processing is used to improve the quality of data set. Feature selection is a technique which is used to remove irrelevant features from the original data set. Feature selection is also known as variable subset selection of relevant features for the use of model construction. Feature selection is help full at the time of analysis of model. Choosing a subset of good features removes the irrelevant data, increases the learning accuracy and improves the comprehensibility. Good feature subset selection algorithms are necessary for machine learning applications. The organization of this paper is as follows: The section 2 provides

detailed description of feature selection. Section 3 gives the different filtering methods. Related work is discussed in section 4. Section 5 gives the Methodology of proposed model. Experimental analysis is shown in section 6. Finally section 7 gives the conclusion.

2. FEATURE SELECTION

In order to make IDS more efficient simplification of feature is done by reducing the data dimension and complexity. The feature selection techniques are mainly divided into two categories, filter and wrapper. The feature selection techniques are mainly divided into two categories, filter and wrapper. Filter method operates without engaging any information of induction algorithm. By using some prior knowledge such as feature should have strong correlation with the target class or feature should be uncorrelated to each other, filter method selects the best subset of features Alternatively, wrapper method employs a predetermined induction algorithm to find a subset of features with the highest evaluation by searching through the space of feature subsets and evaluating quality of selected features. The process of feature selection acts like "wrapped around" an induction algorithm since wrapper approach includes a specific induction algorithm to optimize feature selection; it often provides a better classification accuracy result than that of filter approach. However, wrapper method is more time consuming than filter method due to it is strongly coupled with an induction algorithm with repeatedly calling the algorithm to evaluate the performance of each subset of features. It thus becomes unpractical to apply a wrapper method to select features from a large data set that contains numerous features and instances [12].

3. FILTERING METHODS

There are seven different filter based ranking methods are there, among those six methods are commonly used. The commonly used six methods are chi-squared statistic ($\chi 2$), Information Gain (IG), Gain Ratio (GR), two versions of Relief (RF and RFW) and Symmetric Uncertainty (SU), while the last, Signal-to-noise (S2N) which is less in known. $\chi 2$, IG, GR,RF, RFW and SU are available in the Weka data mining tool[4].

3.1 Chi-square based Attribute Selection

Chi-square test is commonly used method, which evaluates features individually by measuring chi-square statistic with respect to the classes. The statistic is

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{(A_{ij} - E_{ij})}{E_{ij}}$$

Where, k = No. of attributes, n = No. of classes, Aij = number of instances with value i for attribute and j for the class, Eij = the expected No. of instances for Aij. The larger value of the x^2, indicates highly predictive to the class

3.2 Information Gain

Information Gain (IG) is a commonly used measure in the fields of information theory and machine learning. IG measures the number of bits of information gained about the class prediction when using a given feature to assist that prediction. For each feature, a score is obtained based on how much more information about the class is gained when using that feature. The information gain of feature X is shown below,

$$IG(X) = H(Y) - H(Y|X)$$

Where H(Y) and H(Y |X) are the entropy of Y and the conditional entropy of Y given X, respectively. The level of a feature's significance is thus determined by how great is the decrease in entropy of the class when considered with the corresponding feature individually.

A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

3.3 Gain Ratio Attribute Evaluation:

It uses an extension to the information gain uses the gain ratio [5] Gain Ratio (A) = Gain (A)/Split Info (A)

$$\text{Gain Ratio (A)} = \text{Gain (A)/Split Info (A)}$$

This value represents the potential information generated by splitting the training data set.

3.4 Correlation Attribute Evaluation:

Correlation specifies dependence of feature on each other. It represents the linear relationship between the variables or features.

$$\gamma_{A,B} = \sum_{i=1}^{N} \frac{(a_{i-\bar{A}})(b_{i-B})}{N\sigma_A\sigma_B} = \sum_{i=1}^{N} \frac{(a_i b_{i-N\bar{A}B})}{N\sigma_A\sigma_B}$$

Here *N* is the number of tuples, and is the respective values of *A* and *B* in tuple *i*, A

and B are the respective mean values of *A* and *B*, σ*A* and σ*B* are the respective standard deviations of A and B. The value of rA,B lies between -1 and 1. If A and B are completely correlated, rA,B takes the value of 1, if A and B are inversely correlated then rA,B takes value of -1and if A and B are totally independent then rA,B is zero.

3.5 Symmetrical Uncertainty

Symmetric Uncertainty (SU) is a correlation measure between the features and the class and it is obtained below equation,

$$SU = \frac{H(X) + H(Y) - H(X/Y)}{H(X) + H(Y)}$$

where H(X) and H(Y) are the entropies based on the probability associated with each feature and class value respectively and H(X,Y), the joint probabilities of all combinations of values of X and Y .

4. RELATED WORK

Feature subset selection can be viewed as the method of identifying and removing a lot of unrelated and unnecessary features. The reason is , immaterial features do not give the predictive correctness and unnecessary features do not redound to receiving a superior predictor for that they give main data which is previously there in additional feature(s). There are numerous feature subset selection algorithms , a few can successfully remove immaterial features but not succeed to hold unnecessary features however a few of others can remove the immaterial while taking concern of the unnecessary features.

Feature ranking in IDS using combination of filtering methods [2], Zahar et al. , proposed a hybrid feature selection method used to select and rank reliable feature and eliminated irrelevant and useless feature to have more accuracy and reliable intrusion detection process. They have considered a low cost and low accuracy of filtering methods to balance between them. First they have selected two subset of reliable features are created by application of information gain and symmetrical uncertainty filtering methods. In the second phase, the two subsets are merged, weighted and ranked to extract the most important features. This feature ranking which is done by the combination of two filtering methods, leads to higher the accuracy of intrusion detection and they have used KDD99 data set.

Feature Selection for Intrusion Detection using NSL-KDD [3], Hee-Su Chae et.al., and team proposed a performance of standard feature selection methods like CFS(Correlation-base Feature Selection), IG(Information Gain) and GR(Gain Ratio) and to evaluate the classification performance on each of these feature sets using J48 decision tree classifier with full training set and 10-fold cross validation for the testing purposes. In 10-fold cross-validation, the available data is randomly divided into 10

disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining nine sets are used for building the classifier. The test set is then used to estimate the accuracy, and the accuracy estimate is the mean of the estimates for each of the classifiers. Cross-validation has been tested extensively and has generally been found to work well when sufficient data is available.

4.1 Data set

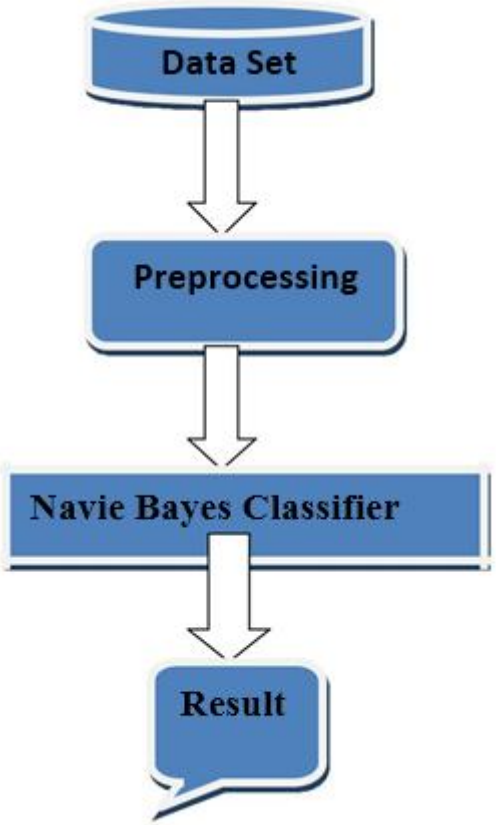
The right selection of data set plays a major role in performance of intrusion detection system. KDD-Cup is the widely used dataset for training and testing of IDS. There are 41 features which are classified into Basic, Content and Traffic features. KDD-Cup is developed on the basis of DARPA'98 data. NSL-KDD is an advanced version of KDD-Cup dataset and doesn't suffer from the shortcomings in KDD-Cup. The following are unique features for which we preferred NSL-KDD over KDD-Cup[6].

- No redundancy of record
- No duplicate
- Less complex level
- Reasonable records

The dataset has about 4,90,000 single connection records with no redundancy. Each connection record has 41 attributes and one class attributes. Class attributes labels connection as normal or attack with exactly one specific attack type. But for analysis researches are using 20% of NSL-KDD dataset that is about 22,495 records with class attribute as a normal or one specific attack type. The attack type includes: DOS, Probe, R2L and U2R attacks[7].

5. METHODOLOGY

Figure 1 shows the general methodology used for an intrusion detection system. Firstly, classify the type of attacks and save the dataset into an ARFF file format. In this proposed work performance is analyzed by using data mining tool called WEKA which is going support the files easily if it in the flat file format (ARFF). Different attribute selection algorithms are applied on the dataset to select the attributes. These selected attributes are used to classify the attribute selection algorithm using Naive Bayes classifier.



z

Figure 1: General Methodology

5.1 NAIVE BAYES

This classifier is based on the elementary Bayes' Theorem. It can achieve relatively good performance on classification tasks. Naive Bayes classifier greatly simplifies learning by assuming that features are independent given the class variable. More formally, this classifier is defined by equation given below [10].

Where X = ( ) denotes a feature vector and Cj, j = 1, 2, ..., N, denote possible Class labels. The training phase for learning a classifier consists of estimating conditional probabilities P (XJ | Ci) and prior probabilities. Here, are estimated by counting the training examples that fall into class ci and then dividing the resulting count by the size of the training set. Similarly, conditional probabilities are estimated by simply observing the frequency distribution of feature XJ within the training subset that is labeled as class ci. To classify a class-unknown test vector, the posterior probability of each class is calculated, given the feature values present in the test vector; and the test vector is assigned to the class that is of the highest probability.

In the experiments, the standard measurements such as detection rate (DR), false positive rate (FPR) were used for evaluation of the performance of intrusion detection tasks. The denotations of True Positive (TP), True Negatives (TN), False Positive (FP), and False Negative (FN) are defined as follows. Equations [10].

True Positive (TP): The number of malicious records that are correctly identified.

True Negatives (TN): The number of legitimate records that are correctly classified.

False Positive (FP): The number of records that were incorrectly identified as attacks however in fact they are legitimate activities.

False Negative (FN): The number of records that were incorrectly classified as legitimate activities however in fact they are malicious.

Detection rate and false positive rate can be detected by using the fallowing equations:

6. EXPERIMENTS AND RESULT

6.1 Setup

Experiments are performed in WEKA3.6 environment using 20% of NSL-KDD dataset on a standalone machine having core-i5 processor with 8GB RAM. WEKA supports many different standard data mining tasks such as preprocessing, classification, clustering, feature selection, regression and visualization. WEKA operates on the predication that the user data is available as a flat file [11].

From the feature selection method applied on NSL-KDD training data set some of features play less role which is very negligible and some play no role. The feature which play no role and play less role can be removed from the list of attributes. The detailed analysis of removed features and the classification result for correctly classified and incorrectly classified instances is shown in the table 2 and table 3 respectively.

By considering 10% of NSL-KDD dataset the number of instances selected is 12598 with all 42 attributes. Number of attacks categorized is shown in Table 1,

Table 1:- categorization of number of instances.

Number of instances	Normal	DOS	Probe	R2L	U2R
12598	6734	4593	1167	98	6

Table 2:- Experimental results with selected features to be removed.

Chi-Square	IG	GR	RF	SU
9,15,20,21	7,9,15,20,21	9,15,20,21	6,7,9,11,15,16,18.19,20,21	7,9,15,20,21

On the basis of analysis of feature selection, we can reduce NSL-KDD data set. The dimension of the data set can be produced by removing the feature numbers given in the table 2[13].

Table 3:- Experimental results for Navie Bayes classifier to find correctly and incorrectly classified instances.

Filtering Methods	Chi-Square	IG	GR	RF	SU
Correctly classified	85.91%	85.92%	86.55%	85.24%	85.93%
Incorrectly classified	14.08%	14.07%	13.45%	14.76%	14.06%

7. CONCLUSION

In this paper, we have proposed a model which is going to suggest that in an intrusion detection system it is not necessary to perform test on all the 41 attributes of NSL-KDD data set. Initially by using a feature selection methods select the features which are necessary. Then, apply the Naïve Bayes machine learning algorithm classify the correctly and incorrectly classified instances.

REFERENCES

[1] Hafiz Muhammad Imran, Azween Bin Abdullah, Muhammad Hussain, Sellappan Palaniappan, Iftikhar Ahmad "Intrusions Detection based on Optimum Features Subset and Efficient Dataset Selection" *IJEIT*, Volume 2, Issue 6, 2012.

[2] Dr. S.Siva Sathya, Dr. R.Geetha Ramani, K.Sivaselvi "Discriminant Analysis based Feature Selection in KDD Intrusion Dataset", *IJCA*, Volume 13, No-11, 2011.

[3] Khedkar S.A., Bainwad A. M., Chitnis P. O. "A Survey on Clustered Feature Selection Algorithms for High Dimensional Data*", IJCIT*, Volume 5, 2014.

[4] Prof. N.S. Chandolikar , Prof. (Dr.) V.D. Nandavadekar "Selection of Relevant Feature for Intrusion Attack Classification by Analyzing KDD Cup 99" *MITIJCSIT*, Volume 2, No-2, 2012.

[5] Zahra Karimi, Mohammad Mansour , Ali Harounabadi "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods", *IJCA*, volume 4, No-78,2013.

[6] K. Khor, C. Ting and Somnuk. "A Feature Selection Approach for Network Intrusion Detection System." *IEEE,* 2009.

[7] Ms Pooja Bhoria, Dr. Kanwal Garg, "Determining feature set of DOS attack", *IJARCSSE*, Volume 3, Issue 5, 2013.

[8] C. Lima, M. Assis and C. Protásio de Souza. "An Empirical Investigation of Attribute Selection Techniques based on Shannon, Rényi and Tsallis Entropies for Network Intrusion Detection." *American Journal of Intelligent Systems* 2.5,2012.

[9] Duch, W. "Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Biesiada, J., and Correlation-Based Filter Solution", 4th International Conference on Computer Recognition Systems, 2005.

[10] Witten, I.H., and Frank, E.. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems). 2005.

[11] Vipin Kumar, Himadri Chauhan, Dheeraj Panwar," K-Means of Clustring Approach to Analyze NSL-KDD Intrusion Detection Dataset" *IJSCE*, Volume 3, Issue 4, 2013.

[12] Theyazn H Aldhyani, Manish R Joshi, "Analysis of Dimentionality Reduction in Intrusion Detection", *IJCII*, volume 4, No-3, 2014.

[13] Karan Bajaj, Amit Arora," Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach", *IJCSI*, Volume 10, Issue 4, 2013.