ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Comparative Analysis of Feature Selection Algorithms

Hancel Anacletus D'Souza, Mani Bushan D'Souza

M.Sc Software Technology, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

Assistant Professor, Dept. of M.C.A., AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

ABSTRACT: The amount of high-dimensional data has increased in past few years. The irrelevant attributes should be removed in order to increase the performance of a system. In Network traffic dataset all the attributes may not contribute in intrusion detection. Various feature selection algorithms are proposed in literature. In this paper a comparative analysis of different feature selection algorithms is done using KDDCUP' 99 data set. Correlation feature selection, Chi Squared attribute evaluation, Consistency subset evaluation, filtered attribute evaluation, filtered subset evaluation, gain ratio attribute evaluation, information gain attribute evaluation, One RA attribute evaluation, Symmetrical uncert attribute evaluation are tested on classifiers Naïve Bayes by using WEKA Tool.

KEYWORDS: Feature Selection; KDDCup99 dataset;

I. INTRODUCTION

In past years there is an increased growth in computer networks. A lot of data is stored over networks by many organizations. It is posinga challenge to detect intrusions since there is huge network traffic. Security threats cause a lot of information security issues.Intrusion Detection System concept was conceived by James Anderson.It identifies malicious activities and alerts network administrator acknowledging about intrusions.

Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. [3]Feature selection has been an active research area in pattern recognition, statistics and data mining communities. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

The KDD CUP 1999 [7] benchmark datasets are used in order to evaluate different feature selection method for Intrusion detection system. 4,940,000 records are available in the data set. Each connection has a label of either normal or the attack type, with exactly one specific attack type falls into one of the four attacks categories [8] as: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack.

II. FEATURE SELECTION METHODS

Feature Selection is also known as variable selection, attribute selection, or variable subset selection in machine learning. Feature Selection algorithms can be divided into three categories they are filter method, wrapper method and hybrid method.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

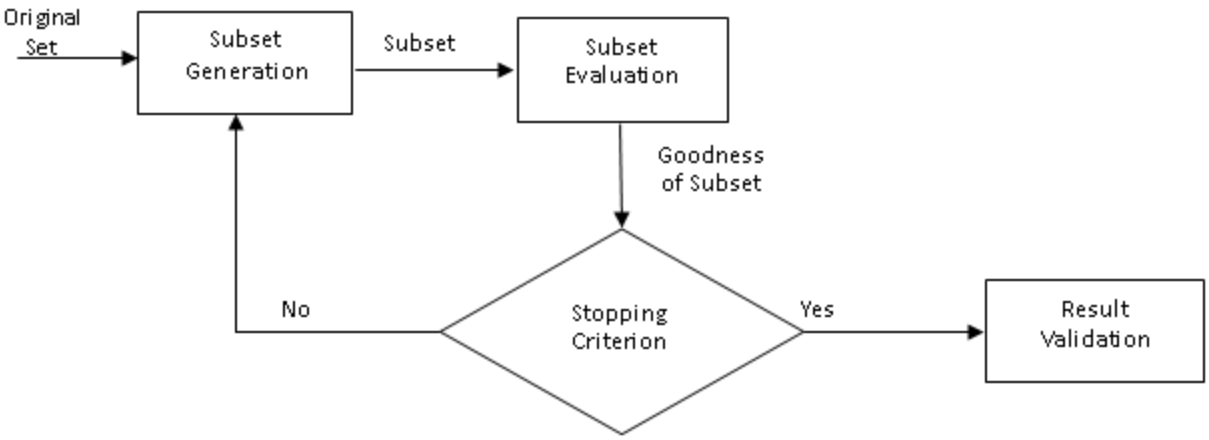

Figure 1. Four key steps for the feature selection process [4]

Subset generation is the process of the search. The process of subset generation has two basic issues to determine a feature subset, namely *search organization* and *successor generation.*

Search Organization

There are three types of search 1) Sequential search 2) Exponential Search 3) Random Search

- Sequential Search

[5]In sequential search, search selects only one among all successors. [5]It is done in an iterative way and the number of possible steps is O(N). It's very easy to implement this method.

- Exponential Search

This search offers best solution.Different heuristic functions can be used to reduce the search space without tempering the optimal solution.BRANCH AND BOUND [6] and Beam Search [9] are evaluated for smaller numbers of subsets for an optimal subset. The search space order is $O(N^2)$.

- Random Search

[5]Random search starts with randomly selected subset. There are two ways to proceed to get an optimal subset. One, generation of the next subset is completely random manner known as the *Las Vegas algorithm*[1].The other is sequential search, which includes randomness in the above sequential approach. The concept of randomness is to avoid local optima in the search space. Search space order is $O(N^2)$.

Evaluation of Subset

The goods of the newly generated feature subset must be evaluated using certain evaluation criteria. An optimal feature subset generated by one criterion may not be same according to the other evaluation criteria. There are two broadly used evaluation criteria, based on their dependency and independence on the algorithms, which are mentioned below.

Independent Criteria

Basically, a filter model is used for independent criteria feature subset selection. It does not involve any learning algorithm. It exploits the essential characteristics of the training data to evaluate the goodness of the feature subset.

General approach for feature selection.

The filter approach:

The filter approach incorporates an independent measure for evaluating features subsets without involving a learning algorithm. This approach is efficient and fast to compute (computationally efficient). However, filter methods can miss features that are not useful by themselves but can be very useful when combined with others. The graphical representation of the filter model is shown in Figure 2.

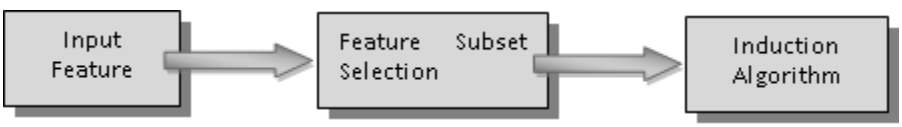

Figure 2 The feature filter model [2]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

The wrapper approach:
Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. The two main disadvantages of these methods are 1) The increasing overfitting risk when the number of observations is insufficient.2) The significant computation time when the number of variables is large.The wrapper approach uses a learning algorithm for subset evaluation. A graphical representation of the wrapper model is shown in Figure 3.

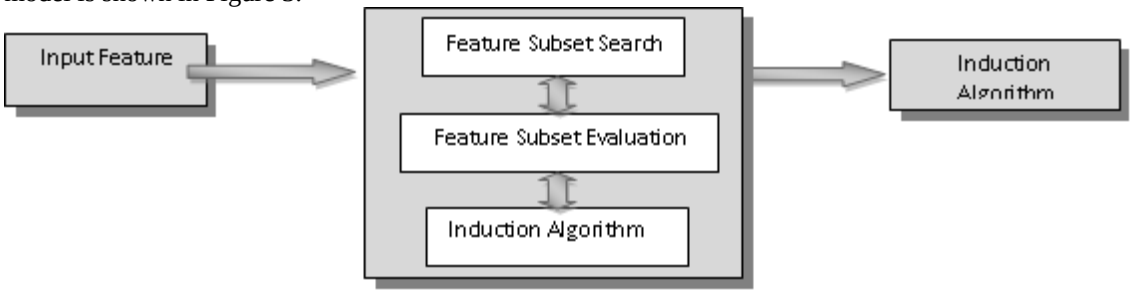

Figure 3 The wrapper model [2]

Embedded Approach
This approach interacts with learning algorithm at a lower computational cost than the wrapper approach. It also captures feature dependencies. It considers not only relations between one input features and the output feature, but also searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subsets for a known cardinality. And then, the learning algorithm is used to select the final optimal subset among the optimal subsets across different cardinality.

III. RESEARCH METHODOLOGY

The aim of this study is to analyse the effect of number of attributes on accuracy of the classifier to classify the instances of network traffic dataset. Maximum accuracy with minimum number of features is required to minimize the computational time of IDS in threat detection. In this study, nine feature selection techniques have been considered to reduce the dimensionality of data. Three classifiers are considered to analyse the effect on accuracy. The methodology adopted for this study is described in figure 4. Data mining tool, Weka, is used to carry out the study Generate all the possible routes.

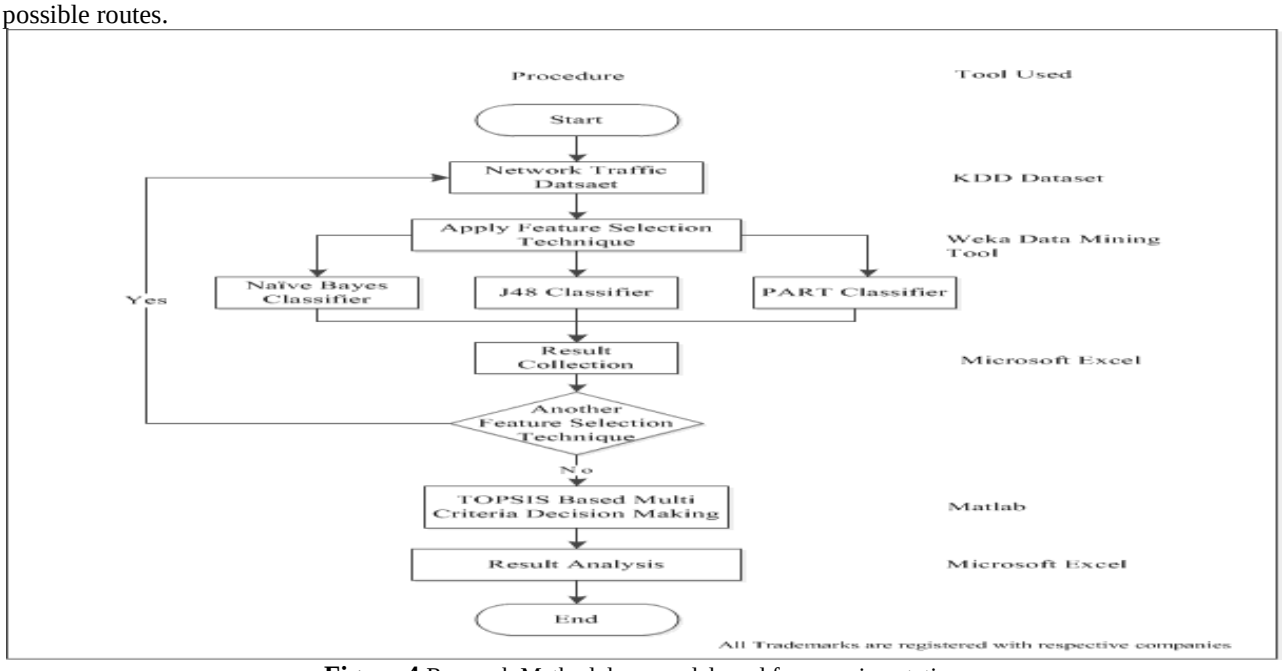

Figure 4 Research Methodology model used for experimentation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

IV. RESULTS

Various parameters used for analysis includes accuracy, number of features used, true positive rate and false positive rate. Table 1 shows the values of different parameters obtained with respect to each feature selection algorithm with Naïve Bayes classifier. If dataset with full feature is taken classifier gives accuracy of 93.565 % but since the number of features are more, computational complexity is high. It can be analyzed that however number of features has some effect on accuracy but it considerably reduce the computational time. If the accuracy criteria is taken then filtered attribute eval, info gain attribute eval, One RA attribute eval and Symmetrical uncert attribute eval algorithm performs very well in case with Naïve Bayes classifier. These algorithms obtained accuracy very close to accuracy of full features dataset but with only 30 features.

PERFORMANCE OF NAÏVE BAYES CLASSIFIER WITH DIFFERENT FEATURE SELECTION TECHNIQUES

Table 1

Feature Selection	Accuracy (%)	No. of Features	TP Rate	FP Rate
Full Features	93.565	41	0.936	0.002
CFS Subset Eval	92.742	10	0.927	0.001
Chi Squared Attribute Eval	93.209	30	0.932	0.002
Consistency Subset Eval	92.317	14	0.923	0.002
Filtered Attribute Eval	93.492	30	0.935	0.002
Filtered Subset Eval	92.715	7	0.927	0.001
Gain Ratio Attribute Eval	89.037	30	0.89	0.001
Info Gain Attribute Eval	93.492	30	0.935	0.002
One Ra Attribute Eval	93.492	30	0.935	0.002
Symmetrical Uncert Attribute Eval	93.492	30	0.935	0.002

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

CFS Subset Eval	97.036	10	0.97	0.002
Chi Squared Attribute Eval	97.534	14	0.975	0.003
Filtered Attribute Eval	97.552	30	0.976	0.003
Filtered Subset Eval	97.026	7	0.97	0.002
Gain Ratio Attribute Eval	97.478	30	0.975	0.003
Info Gain Attribute Eval	97.552	30	0.976	0.003
One Ra Attribute Eval	97.552	30	0.976	0.003
Symmetrical Uncert Attribute Eval	97.552	30	0.976	0.003

V. CONCLUSION AND FUTURE WORK

Network traffic dataset is huge. It may contain millions of instances with hundreds of features. For any IDS it may be impossible to process each instances with all features and attributes. Some features may not be relevant while others may be redundant or have no information regarding intrusion detection. But all these features increase computational time and complexity of IDS. This un-relevant and redundant feature should be discarded so that IDS detects threats in reasonable time. These papers analyze the performance ofclassifier and feature selection techniques. Various parameters like accuracy, number of features, true positive rate and false positive rate are taken for the study. On the basis of accuracy obtained and number of feature suggested by various feature selection techniques, comparative study is carried put.

REFERENCES

1. G. Brassard, P. Bratley, "Fundamentals of Algorithms," *Prentice Hall*, New Jersey, 1996.
2. G. H. John, R. Kohavi, K. Pfleger, "Irrelevant feature and the subset selection problem," in *Proc. of the Eleventh International Conference on Machine Learning*, pp. 121-129, 1994.
3. YongSeog Kim, W. Nick Street, and Filippo Menczer, Feature Selection in Data Mining.
4. M. Dash, H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, Elsevier, pp. 131-156, 1997.
5. Vipin Kumar ,Sonajharia Minz, "Feature Selection: A literature Review",Smart Computing Review, vol. 4, no. 3, June 2014
6. P. M. Narendra, K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transaction on Computer*, vol. 26, no. 9, pp. 917-922, 1977.
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. J. Doak, "An evaluation of feature selection methods and their application to computer security," *Technical report*, Davis CA: University of California, Department of Computer Science, 1992.