A Review on Feature Selection Approaches for Effective Intrusion Detection System

Shruti Dubb¹, Asst. Prof. Yamini Sood²
¹Department of Computer Science, Sri Sai University, Palampur
²Department of Computer Science, Sri Sai University, Palampur

ABSTRACT

Intrusion detection system (IDS) is the system which identifies malicious activity on the network. As the Internet volume is expanding quickly, security against the constant assaults and their quick discovery issues pick up consideration of numerous specialists. Information mining techniques can be adequately connected to IDS to handle the issues of IDS. It can lessen the time intricacy by selecting just valuable components to build model for IDS. There are numerous strategies that are created to choose valuable components. One of the transformative methodologies for valuable choice is hereditary calculation or Genetic Algorithm which is utilized as a hunt strategy while selecting highlights from full NSL KDD information set. The aftereffects of previous approaches when analyzed utilizing classifiers, it indicates enormous development in exactness of a Naïve Bayes classifier with diminished time and least number of highlights.

Keywords - Feature Selection, Genetic Algorithm (GA), Intrusion detection system (IDS), Knowledge Discovery Data mining Dataset (KDD), Naïve Bayes

1. INTRODUCTION

With the tremendous development in the utilization of computer systems step by step, system and additionally data security turns into the prime vital element. The fundamental point of security is to create defensive programming framework which can give three essential security objectives that are authentication, confidentiality and integrity. Intrusion is any movement which tries to disregard these security objectives [1]. The intrusion detection system (IDS) assumes a key part in distinguishing such activities. The term IDS was initially presented by Anderson in 1980. Conventional IDS has huge downsides as far as taking care of high dimensional information. There is a problem of false positives and false negatives in IDS. False positives are the alarms which recognize normal behavior as malicious and false negatives are the alarms which recognize attacks as normal process. [2] A considerable measure of exploration has been going on planned to maintain a strategic distance from them. Information mining is an innovation which utilizes exceptionally created and complex calculations for handling expansive volume of information. [3] Because of this different information mining systems are connected to IDS for enhancing precision. Information sets with several characteristics are pointed as high dimensional information. Addition of new elements more than any point leads to degradation of performance. This is known as the destruction of dimensionality. To defeat this issue one can utilize less number of components than initially to decrease the measurements of the dataset.[4] Selecting features is one of the procedures of dimensionality lessening, in which just restricted components are chosen utilizing an algorithmic methodology disregarding unimportant elements. By methodologies Intrusion Detection System (IDS) is the framework which recognizes offensive action on the system. Information mining strategies can be adequately connected to IDS to handle the issues high dimensional data and to enhance IDS execution. It can decrease the time and increase the quality by selecting just helpful elements to manufacture model for characterization. There are numerous feature selection systems that are created either to choose the elements or concentrate highlights. [5] A developmental methodology for highlight determination is proposed which depends on numerical crossing point rule. Hereditary calculation or Genetic Algorithm is utilized as an inquiry strategy while selecting features from full NSL KDD information set alongside the convergence standard of selecting those just who shows up all around in the investigation. The consequences of methodology when analyzed utilizing classifiers, it indicates huge development in exactness of a Naïve Bayes classifier with decreased time and least number of elements.
2. LITERATURE REVIEW

An intrusion detection system (IDS) can be host based or network based. In this paper network based IDS system is focused to improve its performance.

2.1 Feature Selection (FS)
It is the procedure of selecting a subset of the accessible components to decrease dimensionality of the dataset [6] [7]. In FS repetitive and insignificant elements are disposed off. FS is a powerful machine learning approach which assists helps in building proficient characterization framework. With lessened element subset, the time multifaceted nature is diminished with moved forward precision, of a classifier [8]. There are three standard approaches for highlight choice: wrapper, filter and embedded. In embedded method FS happens as a part of data mining calculation. Channel strategy chooses highlights free of the classifier utilized while as a part of wrapper strategy components are chosen particularly to classifier planned. Filter strategy utilizes any measurable approach to while selecting highlights though wrapper utilizes a learning calculation to locate the best subset of highlights. Computationally wrapper methodology is more costly and slower than the channel approach yet gives more precise results than channel. The FS calculation comprises of two fundamental parts: evaluation functions and search algorithm. Evaluation function is explained which approach is used for selection. According to their working, search methods can be classified as exponential, sequential or randomized. The exponent method has exponential complexity, randomized selects the features randomly giving high accuracies and in sequential method features are linearly added or subtracted.

2.2 Correlation based Feature Selection (CFS)
Correlation based FS determines the value of a subset of features utilizing heuristic strategies. Feature which is very connected with a class is considered as great and chose. In every subset quality are chosen by considering the level of excess in the middle of them and prescient capacity of every individual component [9]. Thus, there is a need to characterize a proper relationship measure which can list most critical and profoundly viable highlights.

2.3 Information Gain (IG)
Information Gain calculates the entropy value (i.e. how much information it is giving), for each feature. Entropy is a measure of the uncertainty associated with a random variable. With the help of this value we can determine most useful feature for classification. Higher the entropy value, the feature contains more information [10].

2.4 Correlation Attribute Evaluator (CAE)
Attribute redundancy can be evaluated through correlation analysis. The correlation between two attributes, X and Y, can be evaluated by finding correlation coefficient. A good feature is that which is having a higher correlation coefficient between it and class. In correlation attribute evaluator method the attributes are considered based on their values where each value is treated as an indicator. Attributes should be of nominal type as input for evaluation. CAE uses Pearson’s formula for computing correlation coefficient.

2.5 Genetic Algorithms (GA)
Genetic algorithms (GA) are an adaptive heuristic search method based on the idea of natural selection [11]. They are inspired by Darwin’s theory of evolution – “survival of the fittest”, which is one of the randomized search techniques. The algorithm begins with a set of individuals (chromosomes) called as population. Individual chromosome consists of a set of genes that could be bits, numbers or characters. Individuals are selected according to their fitness value for reproduction. Higher the fitness value more is the chances of an individual being selected.

Crossover and mutation is responsible for producing new population. Crossover accelerates the search early in the evolution of the population, while mutation is
responsible for restoring the lost information to population by local or global movement in the search space. The process is iteratively repeated several times until stopping criteria are met or optimal solution is reached.

3. CONCLUSION

According to previous approach, numerical convergence guideline based imaginative methodology utilizing Genetic Algorithm (GA) for feature selection. Feature selection is done utilizing distinctive feature selection (FS) techniques like CFS, IG and CAE and their impact on the execution of two ordinarily utilized classifiers, Naive Bayes and J48, is tried. From the exploratory results it can be inferred that the strategy helps in selecting the minimum number of elements or features from the NSL KDD information set which enhances the Naïve Bayes classifier precision alongside reduced time multifaceted nature.

REFERENCES