Nallamala Srilatha* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.9, Issue No.3, April – May 2021, 9949 - 9951.

Building A Malware Finding System Using A Filter-Based Feature Selection Algorithm

Ms. NALLAMALA SRILATHA
M.Tech Student,Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology, Bahadurpally, Hyderabad, India.

Mrs. MANTRI GAYATRI
Associate Professor, Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology, Bahadurpally, Hyderabad, India.

Abstract: **Flexible Mutual Information Feature Selection is another supervised filter-based feature selection formula that has recently been proposed. With FMIFS, there's no doubt about it, MIFS and MMIFS are outdated. According to FMIFS, a revision to Battiti's formula would help cut down on redundancy among features. Redundancy parameters are no longer required in MIFS and MMIFS because of FMIFS. MIFS and MMIFS are unquestionably better alternatives to FMIFS. Based on the advice of FMIFS, Battiti's formula should be updated to minimize redundancy. In FMIFS, the redundant parameter is eliminated and it results in MIFS and MMIFS. None of the existing technologies are capable of fully safeguarding the internet software and operating networks against threats like DoS attacks, spyware, and adware. Incredible amounts of network traffic pose a major obstacle to IDSs. Our function selection formula contributed significantly more important functionality to LSSVM-IDS in regards to improving LSSVM-IDS' accuracy while minimizing the use of computation in comparison to other approaches. This feature selection method is especially suitable for features that are dependent on either a linear or nonlinear relationship. To provide accurate classification, we have provided a formula based on mutual knowledge, which mathematically selects the perfect function. Its utility is measured by taking into account the use of network intrusion detection. Data with redundant and irrelevant functionality has created a long-term traffic condition. It not only slows the overall classification process, but it also impedes classifiers from making correct decisions, specifically when handling large amounts of data.**

Keywords: **Linear Correlation Coefficient; Intrusion Detection; Mutual Information;**

I.INTRODUCTION:

Applying protection strategies that are both proactive and scalable is now much more important than it was in the past. A more complete protection is provided by the lines that have been blended together. Invasion recognition system is one such kind of protection prevention, for example. When it comes to massive databases, the features they contain can be noisy, repetitive, or lacking in information, and therefore present significant obstacles to discovering and knowledge modeling. In a paper written by Mukkamala et al., numerous approaches to detect intrusions were studied, including Artificial Neural Systems, SVMs, and Multivariate Adaptive Regression Splines. Toosi et al. used a genetic formula to refine the architectures of neuron-fuzzy classifiers, which were used alongside other recognition mechanisms. Computational uncertainty results from the large amount of data that must be categorized. To handle the feature selection issues we've proposed a mix of different formulas, a combination of various feature selection methods [1]. This modern feature selection approach seeks to implement theoretical study of reciprocal data to evaluate feature-to-output class dependency. In order to consider multiclass classification issues, we develop our suggested structure. This also is to show the fruitfulness and usefulness of the technique proposed? Since the feature selection process has no free parameters, it is an extension of Mutual Information Feature Selection and Modified Mutual Information-based Feature Selection.

LiteratureSurvey:Function selection techniques are commonly broken down into filter and wrapper techniques. With regard to filter methods, on the other hand, wrapper methods are often a lot more computationally expensive when processing large-scale data or handling high-dimensional data. The feature selection formula suggested by Mukkamala and Sang in KDD Cup 99 dataset minimizes the feature space of the dataset. In hierarchical clustering, the algorithm for assigning low-quality training data was used to reduce the amount of training data used during training, thus increasing the quality of training data used during testing to produce a better-performing classifier. The right collection of features was adopted by the LS-SVM classifier to be used for training, and for creating the IDS.

II. PROBLEM STATEMENT:

Intelligent intrusion prevention techniques have been developed for the purpose of securing the network. One of the earliest attempts to develop intrusion detection systems was bagged boosting, using C5 decision trees and Kernel Miner. The researchers in this study wanted to find out whether it was possible to use Artificial Neural Networks (ANN), Support Vector Machines (SVMs) and

Multivariate Adaptive Regression Splines (MARS) to detect network intrusions. While current approaches remain incapable of entirely defending websites and data networks against the growing number of sophisticated cyber attack tactics, including DoS attacks and computer ransom ware, they can be improved upon. These large scale network traffic records pose a significant obstacle to IDSs. An increase in "big data" makes identification more complex, causing slower and less accurate classifications. Problems in mathematical classification are typically accompanied by increases in numerical complexity. The difficulty of working with large-scale datasets is due to the presence of repetitive, uninformative, or inaccurate attributes that make finding information or formulating models difficult.

III. PRAPOSED METHODOLOGIES:

Involving a decision tree and support vector machine (SVM) and we have given the details for implementation (HFSA). In Step I, the HFC designs are developed. This part looks over the initial data to weed out irrelevant and redundant functions. Pre-selected features shrink the spectrum over which the wrapper processes (the lower phase) searches for potential matches (the output of the upper phase). In this article, the major contributions are described as follows [2][3]. A novel feature selection approach is proposed in which theoretical consideration of shared knowledge is used to measure feature dependency and class performance. Classifiers are built around the features that are most appropriate, with some features selected for removal. The feature selection approach suggested does not include any free parameter, such as in MIFS and MMIFS. So as a result, it is free from being negatively affected by assigning value to a free parameter and, hence, can be absolutely assured. Additionally, the proposed approach is effective in a variety of applications and much more robust as compared to the Wrapper-based Feature Selection Algorithm (HFSA) where computationally intensive wrapper-based feature selection is used. Additionally, we carry out comprehensive tests on two widely used IDS databases, as well as the dataset we use. KDD dataset is old and does not include most novel attack patterns, which make IDS evaluation tricky. Moreover, these datasets are widely cited in the literature as evidence of the efficacy of intrusion detection systems (IDS). Data with varying sample sizes and different numbers of features will make feature selection algorithms challenging to evaluate comprehensively [4]. Our approach does not only focus on binary classification issues, but it is geared towards dealing with multiclass classification problems as well. This demonstration will demonstrate the efficacy and practicality of the suggested procedure. The introduction of FMIFS represents an advance over MIFS and MMIFS. In an effort to minimize redundancy, FMIFS proposes a modification to Battiti's algorithm. In MIFS and MMIFS, the redundant parameter is eliminated.

IV. ENHANCED SYSTEM:

Data Preprocessing: The specific features created by processing the data obtained during the data collection phase are referred to as basic features, like the ones found in KDD Cup 99 dataset. For each record in the input data, the qualified classifier requires that it be interpreted as a vector of real numbers. As a result, every dataset features one or more symbolic attributes, and these attributes are first expressed as numbers. An example of a function in the KDD CUP 99 dataset is numerical, as well as symbolic. The protocols such as TCP, UDP, and ICMP and the services like HTTP, FTP, Telnet, and soon have been chosen for the logos as well (e.g., SF, REJ and so on). Instead of assigning numerical values to the categorical properties, the procedure actually substitutes the values [5]. This is a vital part of the continuum of data preparation until all symbolic qualities have been transformed into numerical values. Data normalization is a method of scaling the value of each parameter into a uniform collection, thereby eliminating the inherent bias favoring features with higher values.

Filter based feature selection: Correlations may be used to calculate a statistical measurement of dependency, known as the correlation coefficient. However, given the real-world applications, it is important to take nonlinear relationships into consideration. To say that two variables are not linearly dependent does not mean that a straight line can't calculate their relationship. In order to correctly describe a relationship, no matter whether it is linear or nonlinear, we must use a tool that is capable of performing analysis on all dependent variables. Additionally, this work attempts to find an optimum set of features regardless of how well they correlate [6]. For the feature selection process, we've developed two algorithms. As follows: The method for selecting features may be configured to follow the shared knowledge of the dataset.

Attack classification &Recognition: Designing classifiers to differentiate between two classes is usually easier than thinking in terms of multiclasses when faced with a classification problem. The first judgment boundary in this case is more straightforward because of that. The studies in this paper have two distinct classes, where similar records for the "standard" class are recorded as normal data, whereas those records that don't fit are classified as attack data. There are two widely used approaches for dealing with an issue with more than two classes: OVO and OVA (OVA). The regular and intrusion traffics are found by using the classifier that has been educated using the best

subset of features, which contain the most associated and significant features [7]. The trained model is then used to search for attacks, whereupon the test data is sent to it for inspection. All records that have an identical ID to the predefined class are deemed standard data, while all other records are flagged as an attack. The classifier model determines that the record is abnormal. Because the type of attacks is based on subclass characteristics, the record's type can be deduced from this.

Performance Evaluation: Of the tests conducted in the KDD Cup 99 datasets, the bulk are performed on the IDS systems. Also, these datasets feature distinct data sizes and a wide range of features, offering broad feature validation tests. The KDD Cup 99 dataset is widely applied in order to test the performance of intrusion detection systems. Standard classes, along with four distinct attack groups (i.e., DoS, Probe, U2R and R2L). This sample dataset includes data with about five million connection records, which are training data, and data with about two million connection records, which are testing data. Each dataset includes a number of records that are labeled "natural" or "assault," and it includes quantitative and qualitative features that number in the hundreds.

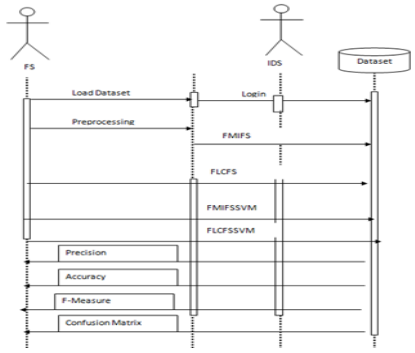

Fig 1: intrusion detection system

Several studies have investigated the proposed LSSVMIDS' efficiency and efficacy. To ensure that these calculations are correct, the following measures are used: Identification rate, false positive rate, F-measure, and accuracy rate.

V. CONCLUSIONS:

A filter-based feature selection algorithm that relies on mutual information feature selection has been proposed, known as Flexible Mutual Information Feature Selection (FMIFS). A big improvement over MIFS and MMIFS is given by FMIFS. In order to improve efficiency, FMIFS proposes a modification to Battiti's algorithm and reduces function redundancy. The redundancy parameter in MIFS and MMIFS is eliminated by FMIFS. Choosing the right value for this parameter does not have a clear protocol or rule. The LSSVM and FMIFS methods are then used to build IDS. The LSSVM operates with equality constraints rather than inequality constraints in order to solve a different classifier formulation, one which asks how many non-zero coefficients the equations must have. This analysis shows that the suggested identification scheme is highly effective at finding intrusions through computer networks. When compared to the other state-of-the-art versions, the overall result has been that LSSVM-IDS and FMIFS performed the best. Since the FMIFS feature selection algorithm has shown promising results, the search strategy can be further optimized to improve the overall efficiency. Additionally, our future experiments should account for the effect of the unbalanced sample distribution on IDS.

REFERENCES:

[1] R. Agarwal, M. V. Joshiy, Pnrule: A new framework for learning classier models in data mining (a case-study in network intrusion detection), Citeseer2000.

[2] D. S. Kim, J. S. Park, Network-based intrusion detection with support vector machines, in: Information Networking, Vol. 2662, Springer, 2003, pp. 747–756.

[3] Mohammed A. Ambuscade, Member, IEEE, Xingjian He*, Senior Member, IEEE,Priyadarsi Nanda, Senior Member, IEEE, and Zhiyuan Tan, Member, IEEE, “Building an intrusion detection system using afilter-based feature selection algorithm”, ieee transactions on computers,2016.

[4] G. Kim, S. Lee, S. Kim, A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, Expert Systems with Applications 41 (4) (2014) 1690–1700.

[5] S.-W. Lin, Z.-J.Lee, S.-C.Chen, T.-Y. Tseng, Parameter determination of support vector machine and feature selection using simulated annealing approach, Applied soft computing 8 (4) (2008) 1505–1512.

[6] Y.-I. Moon, B. Rajagopalan, U. Lall, Estimation of mutual information using kernel density estimators, Physical Review E 52 (3) (1995) 2318–2321.

[7] A. M. Ambusaidi, X. He, P. Nanda, Unsupervised feature selection method for intrusion detection system, in: International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2015.

