**Building an intrusion detection system using a filter-based feature selection algorithm**

**ABSTRACT:**

Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions, especially when coping with big data. In this paper, we propose a mutual information based algorithm that analytically selects the optimal feature for classification. This mutual information based feature selection algorithm can handle linearly and nonlinearly dependent data features. Its effectiveness is evaluated in the cases of network intrusion detection. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is built using the features selected by our proposed feature selection algorithm. The performance of LSSVM-IDS is evaluated using three intrusion detection evaluation datasets, namely KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The evaluation results show that our feature selection algorithm contributes more critical features for LSSVM-IDS to achieve better accuracy and lower computational cost compared with the state-of-the-art methods.

**EXISTING SYSTEM:**

* A significant amount of research has been conducted to develop intelligent intrusion detection techniques, which help achieve better network security. Bagged boosting-based on C5 decision trees and Kernel Miner are two of the earliest attempts to build intrusion detection schemes.
* Mukkamala et al. investigated the possibility of assembling various learning methods, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to detect intrusions.

**DISADVANTAGES OF EXISTING SYSTEM:**

* Existing solutions remain incapable of fully protecting internet applications and computer networks against the threats from ever-advancing cyber attack techniques such as DoS attack and computer malware.
* Current network traffic data, which are often huge in size, present a major challenge to IDSs. These “big data” slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data.
* Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity.
* Large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and data modeling.

**PROPOSED SYSTEM:**

* We have proposed a hybrid feature selection algorithm (HFSA). HFSA consists of two phases.
* The upper phase conducts a preliminary search to eliminate irrelevant and redundancy features from the original data. This helps the wrapper method (the lower phase) to decrease the searching range from the entire original feature space to the pre-selected features (the output of the upper phase). The key contributions of this paper are listed as follows.
* This work proposes a new filter-based feature selection method, in which theoretical analysis of mutual information is introduced to evaluate the dependence between features and output classes.
* The most relevant features are retained and used to construct classifiers for respective classes. As an enhancement of Mutual Information Feature Selection (MIFS) and Modified Mutual Information based Feature Selection (MMIFS), the proposed feature selection method does not have any free parameter, such as in MIFS and MMIFS. Therefore, its performance is free from being influenced by any inappropriate assignment of value to a free parameter and can be guaranteed. Moreover, the proposed method is feasible to work in various domains, and more efficient in comparison with HFSA, where the computationally expensive wrapper-based feature selection mechanism is used.
* We conduct complete experiments on two well known IDS datasets in addition to the dataset used. This is very important in evaluating the performance of IDS since KDD dataset is outdated and does not contain most novel attack patterns in it. In addition, these datasets are frequently used in the literature to evaluate the performance of IDS. Moreover, these datasets have various sample sizes and different numbers of features, so they provide a lot more challenges for comprehensively testing feature selection algorithms.
* Different from the detection framework proposed that designs only for binary classification, we design our proposed framework to consider multiclass classification problems. This is to show the effectiveness and the feasibility of the proposed method.

**ADVANTAGES OF PROPOSED SYSTEM:**

* FMIFS is an improvement over MIFS and MMIFS.
* FMIFS suggests a modification to Battiti’s algorithm to reduce the redundancy among features.
* FMIFS eliminates the redundancy parameter required in MIFS and MMIFS.

**SYSTEM ARCHITECTURE:**



**MODULES:**

* Data Preprocessing
* Filter based feature selection
* Attack classification & Recognition
* Performance Evaluation

**MODULES DESCSRIPTION:**

**Data Preprocessing**

* The data obtained during the phase of data collection are first processed to generate the basic features such as the ones in KDD Cup 99 dataset. The trained classifier requires each record in the input data to be represented as a vector of real number. Thus, every symbolic feature in a dataset is first converted into a numerical value. For example, the KDD CUP 99 dataset contains numerical as well as symbolic features. These symbolic features include the type of protocol (i.e., TCP, UDP and ICMP), service type (e.g., HTTP, FTP, Telnet and so on) and TCP status flag (e.g., SF, REJ and so on). The method simply replaces the values of the categorical attributes with numeric values.
* An essential step of data preprocessing after transferring all symbolic attributes into numerical values is normalisation. Data normalisation is a process of scaling the value of each attribute into a well-proportioned range, so that the bias in favor of features with greater values is eliminated from the dataset.

**Filter based feature selection**

* If one considers correlations between network traffic records to be linear associations, then a linear measure of dependence such as linear correlation coefficient can be used to measure the dependence between two random variables. However, considering the real world communication, the correlation between variables can be nonlinear as well. Apparently, a linear measure cannot reveal the relation between two nonlinearly dependent variables. Thus, we need a measure capable of analysing the relation between two variables no matter whether they are linearly or nonlinearly dependent. For these reasons, this work intends to explore a means of selecting optimal features from a feature space regardless of the type of correlation between them.
* We develop two algorithms for feature selection process. There are: Flexible mutual information based feature selection and Feature Selection Based on Linear Correlation Coefficient

**Attack classification & Recognition**

In general, it is simpler to build a classifier to distinguish between two classes than considering multiclasses in a problem. This is because the decision boundaries in the first case can be simpler. The first part of the experiments in this paper uses two classes, where records matching to the normal class are reported as normal data, otherwise are considered as attacks. However, to deal with a problem having more than two classes, there are two popular techniques: \One-Vs- One” (OVO) and \One-Vs-All” (OVA).

After completing all the aforementioned steps and the classifier is trained using the optimal subset of features which includes the most correlated and important features, the normal and intrusion traffics can be identified by using the saved trained classifier. The test data is then directed to the saved trained model to detect intrusions. Records matching to the normal class are considered as normal data, and the other records are reported as attacks. If the classifier model confirms that the record is abnormal, the subclass of the abnormal record (type of attacks) can be used to determine the record’s type

**Performance Evaluation**

* The majority of the IDS experiments were performed on the KDD Cup 99 datasets. In addition, these datasets have different data sizes and various numbers of features which provide comprehensive tests in validating feature selection methods.
* The KDD Cup 99 dataset is one of the most popular and comprehensive intrusion detection datasets and is widely applied to evaluate the performance of intrusion detection systems. It consists of five different classes, which are normal and four types of attack (i.e., DoS, Probe, U2R and R2L). It contains training data with approximately five million connection records and test data with about two million connection records. Each record in these datasets is labeled as either normal or an attack, and it has 41 different quantitative and qualitative features.
* Several experiments have been conducted to evaluate the performance and effectiveness of the proposed LSSVMIDS. For this purpose, the accuracy rate, detection rate, false positive rate and F-measure metrics are applied.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

* System : Pentium Dual Core.
* Hard Disk : 120 GB.
* Monitor : 15’’ LED
* Input Devices : Keyboard, Mouse
* Ram : 1GB.

**SOFTWARE REQUIREMENTS:**

* Operating system : Windows 7.
* Coding Language : JAVA/J2EE
* Tool : Netbeans 7.2.1

**REFERENCE:**

Mohammed A. Ambusaidi, Member, IEEE, Xiangjian He\*, Senior Member, IEEE, Priyadarsi Nanda, Senior Member, IEEE, and Zhiyuan Tan, Member, IEEE, “Building an intrusion detection system using a filter-based feature selection algorithm”, **IEEE TRANSACTIONS ON COMPUTERS,2016.**