IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

Network Anomaly Intrusion Detection System based on SVM and Gradient Boosted Trees

Brunel Elvire Bouya-Moko[1], Edward Kwadwo Boahen[2]

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China[1,2]

Abstract: Intrusion detection system (IDS) has recently become one of the fundamental parts of the security field. It is mainly comprised of two methods, namely anomaly detection and misuse detection. The focus of this paper is on a Network IDS (NIDS) based on feature selection by combining Support Vector Machine (SVM) and Gradient Boosted Trees algorithms. Different approaches have been used for increasing the accuracy to detect the intrusion. The first approach is the filter method using Fisher score and ReliefF score, the second one is the wrapper method and the third approach which brings novelty to this research is the combination of Fisher score and ReliefF score. However, the analysis of the technique is done using SVM with RBF-Kernel and Gradient Boosted Trees. This paper also includes Cross-Validation folds to perform a 10-flods Cross-Validation method for training and validation.

Keywords*:* Intrusion Detection System, Support Vector Machine, Gradient Boosted Trees, Feature ranking and selection.

I. INTRODUCTION

Network based computer systems are now playing an increasingly important role in the modern society, and have become the target of hackers and criminals. As a result, researchers make every effort to develop innovative security solutions to enable system administrators to protect network-based computer systems. An Intrusion is any group of activities attempting to threaten the integrity, confidentiality or availability of system resources or services [13]. A designed IDS focused on detecting attacks that target computer networks is consider as a Network IDS (NIDS) [9]. IDSs are primarily designed to prevent attacks in network-based computer systems; its goal is to alert system administrators to a variety of suspicious actions and provide prevention advice. IDS should be capable of distinguishing normal and abnormal activities [11]. ID methods are basically grouped into two different categories: misuse detection methods, and anomaly-based techniques [33] [31]. The misuse-based ID approach defines and monitors malicious activities. This method is very effective in detecting attacks. However, it misses original attacks if these new attacks are just small variations with respect to the old ones. Anomaly-based ID techniques, on the other hand, create a system to identify usage in a regular system before observing harmful actions. This methodology is particularly effective in detecting novel threats that misuse-based techniques are unable to detect. [31].

Numerous researches have been done in the IDS field, however SVM and Gradient Boosted Trees are very popular ML algorithms as they have proven successful through various fields [32] [8]. But the training is time consuming in some ML applications such as Bio-Informatics and Data Mining because of the amount of data. Hence, a feature selection step is necessary for reducing the number of input variables to lower the computational cost of modelling and, in some situations, to increase the model's performance. Feature subset selection approaches are generally split into two different types: filter and wrapper approaches [4]. Filter approaches are usually based on statistical or probabilistic database properties and are not dependent on learning machines. Filters are computationally efficient and ideal for high-dimensional databases because they do not use a learning machine. Wrapper models, on the other hand, employ learning machines to choose feature subsets depending on prediction performance. Hence, when compared to filters, their processing cost is relatively large, and they are thus ineffective for large-dimensional databases. Wrappers have a significant advantage over filters in terms of prediction accuracy. The results of a wrapper's search for the best feature subset are often more promising than results based on filters since the wrapper's search for the best feature subset is led by prediction accuracy [27].

The major objective of this research is to devise a new approach to network IDS combining SVM with RBF-Kernel and Gradient Boosted Trees, with the application based on the filter and wrapper methods for performing feature selection. The filter method is applied using fisher score. Afterward, fisher score and reliefF score are combined to rank the features where highly ranked features are selected then send to a predictor. The weighted average of fisher score rank and reliefF score rank are computed for sorting features. Furthermore, cross-validation fold is employed to perform 10-flods cross-validation method to facilitate the model estimation and variable selection. The proposed approach has a reasonable

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

compromise between quality and performance, providing a solution to the limitations inherent to the typical SVM and Gradient Boosted Trees and outperforming other IDS models.

The key part of this research work is given below as follows,

- **Proposed NIDS approach based on SVM and Gradient Boosted Trees:**
  To effectively detect network intrusions, an efficient and robust intrusion detection approach based on the SVM and Gradient Boosted Trees is introduced. Herein, features selection is performed by various approaches, namely filter technique using Fisher score, wrapper technique and the incorporation of Fisher score and ReliefF score.

The remainder of this paper is presented as follow: section 2 focuses on related work of network intrusion detection with respect to machine learning algorithms. Section 3 provides necessary information related to the topic. Section 4 is concerned with the methodology approaches used for this research with all details. Section 5 portrays the experimental results and finally section 6 concludes the paper.

II. MOTIVATION

The existing NIDS approaches are explained in this section along with the advantages and disadvantages.

A. Literature review

The various existing research techniques based on NIDS are illustrated in this section.

Depren, O., *et al.* [6] presented a new IDS model that combines anomaly and misuse detection techniques. To model legit activities, the designed anomaly detection module employs a Self-Organizing Map structure. When an activity deviates significantly from usual behaviors, it is considered as a threat. To classify various sorts of threat, the designed misuse detection module employs the J.48 decision tree method. Other approaches are outperformed by this proposed hybrid model. However, the anomaly module has a higher rate of false positives. Wang, G., *et al.* [31] developed anomaly-based intrusion detection algorithm to detect anomalous connections. In this research work, the feature selection approach is performed based on the SVM (named FS-SVM) for minimizing the dimensionality of sample data. In addition, the authors introduced a multiclass SVM methodology with PSO-optimized parameters for learning a classifier for identifying multiclass attacks (MSVM-PSO). The effectiveness and performance of the proposed methodologies have been successfully tested; nevertheless, the techniques cannot detect abnormal physical or virtual nodes in novel computing environments such as cloud computing. Kim, G., *et al.* [14] devised a hybrid intrusion detection approach for detecting intrusions. This approach hierarchically incorporated a misuse detection technique and an anomaly detection technique in a decomposition arrangement to make IDS systems more effective. The approach needs to be improved in order to uniformly split the normal data into each subset without compromising the misuse detection accuracy, which should result in a large improvement. Varshavsky, R., *et al.* [30] developed a filter method, called novel unsupervised standard using SVD-entropy to choose the features based on the entropy (CE) computed on a leave-one-out basis. Herein, various methods with respect to these criteria are utilized, afterward the usefulness is tested on three other biological benchmarks. All these feature filtering methods outdone other conventional unsupervised filtering approaches.

Maldonado, S., & Weber, R [19] developed a new wrapper algorithm to perform the feature selection using SVM with kernel function. The method is mainly focused on sequential backward selection, with the number of errors in a validation subset serving as the criterion for determining which features will be deleted during each iteration. This developed approach is compared with other existing algorithms such as filter technique or Recursive Feature elimination SVM for demonstrating its efficiency. In regards of the approach's drawbacks, more research is needed in a variety of directions. To begin, the proposed feature selection wrapper methodology can be used with SVM variations such as alternative kernel functions and Support Vector Regression. Additionally, it would be much more accurate to combine the Hold Out SVM and weighted SVM approaches to compensate for the negative impact produced by unbalanced data sets in model creation, which is a common problem. Wrapper approaches generally outdone filter approaches because they determined the feature subset based on the prediction accuracy. For that reason, Hu, Z., *et al.* [10] developed a hybrid filter-wrapper technique for the short-term load (STLF) forecasting. In this technique, the Partial Mutual Information-enabled filter technique is utilized for filtering the inappropriate and unnecessary features, whereas the wrapper technique applied in this algorithm reduced the redundant features without any performance degradation in forecasting the accuracy. The results showed that the suggested hybrid filter–wrapper-based model outperforms other well-known models when it comes to forecasting. However, the authors only evaluated the most popular input possibilities when choosing the model. Maldonado, S.,*et al.* [20] devised a kernel-penalized SVM (KP-SVM) for selecting important features in a simultaneous way during the classifier construction with the penalty of every feature's use in the double SVM formulation. The

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

proposed algorithm is dependent on the non-linear optimization issue, which is computationally treatable. However, it possesses higher computational cost when facing a huge number of input features. Ring, M., & Eskofier, B. M., [21] introduced the exact Gaussian Radial Basis Function (RBF) kernel. Here, a user-defined approximation quality is given in the categorization phase for accelerating the evaluations-based on SVM. The basic principle of the current technique could be used for approximating other kernels in future study.

Liu, Y., & Zheng, Y. F., [18] designed a filtered and supported sequential forward search algorithm (FS_SFS) for SVM. This technique is comprised of two major properties for reducing the computational time. Initially, subset of samples is maintained for training the SVM, then discriminant capability of the individual feature and the correlation are applied for filtering the non-essential features. This paper's contribution is significant because it considerably saves the training time. Tu, I. B [26] devised a new feature selection approach that incorporates feature wrapper and feature filter techniques for identifying important input variables in the systems with continuous domains. This newly designed approach utilized functional dependency notion, correlation coefficients and the K-nearest neighborhood (KNN) technique for implementing the feature filter and feature wrappers. However, for enhancing the method, a more elaborate approach for approximate functional interdependence could be provided. Bhuyan, M. H., *et al* [2] presented a survey of complete outline of different facets of network-driven anomaly detection to help researchers to rapidly familiar with all aspect related to network-based anomaly detection. Chandola, V.,*et al* [3] presented a survey to provide an organized and thorough overview of the research related to anomaly detection. In this study, various existing methods were grouped into dissimilar categories based on latent methods that were adopted by each technique.

III. PROPOSED NIDS METHOD FOR NETWORK ANOMALY DETECTION USING SVM AND GRADIENT BOOSTED TREES

This section illustrates the NIDS technique for network anomaly detection using SVM and Gradient Boosted Trees. The first part is data prepossessing where data are transformed to be used in the whole method. The second part is features selection where the first approach is filter method using Fisher Score and ReliefF score, the second approach is wrapper method and the third approach is the combination of Fisher Score and ReliefF score in the following way: every feature gets its rank according to Fisher Score and rank according to ReliefF score. The weighted average of Fisher Score rank and ReliefF score rank is computed and the features are based on the combined rank. Experiments are conducted with different weights for both Fisher score ranks and ReliefF score ranks. Calculations are conducted on selected features using SVM and Gradient Boosted Trees.

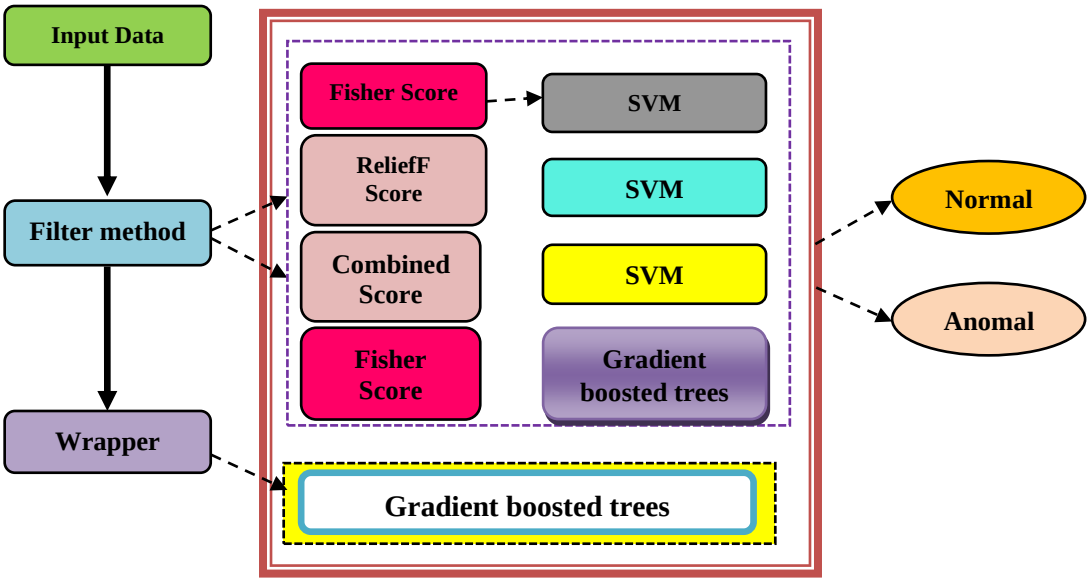


Fig. 1 Schematic view of the developed NIDS technique based on based on SVM and Gradient Boosted Trees

A. Input data acquisition

Let us consider the database as F with x number of network intrusion data D, which is depicted as,

$$F = \{D_1, D_2, ..., D_p, ...D_x\} \quad (1)$$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

where, D represents the intrusion data, F indicates the database, and D_p demonstrates the intrusion data situated at p^{th} index. From the intrusion dataset, intrusion data of input network D_p is considered and is subjected to the feature selection module in order to perform the ID process.

B. Feature selection for SVM

Datasets usually contain huge amount of feature that requires more time to be processed. It is a difficult task because ML algorithms work easily and quickly when facing a small amount of feature. Theoretically, the increase of feature vector size provides additional discriminating power. However, the feature vectors considerably slow down the learning procedure, thereby facilitating the classifier to compromise the generalization performance, and also to over fit the training samples. Hence, the feature selection mechanisms are utilized for minimizing the dataset dimensionality to enhance the learning performance, minimize the cost of computation, and also to simplify the dataset assessment. There are two various techniques employed for selecting the features, such as filter technique, and wrapper technique [16]. This research utilized both the filter and wrapper-based techniques.

1. Filter method:

The filter approach selects features with predictor independence based on the wide-ranging behavior of the training data. It performs the process of feature selection as a preprocessing phase with no use of an induction algorithm [5]. The relief algorithms are known as a single class of filter-style feature selection algorithm [28]. The relief algorithm was originally designed by [15], and the main idea is to evaluate the feature quality with respect to how well the attribute values differentiate among the instances that are closer to each other. Moreover, the fundamental idea of Relief algorithms is described below as follows:

Basic idea of Relief algorithm:

An instance R_i is selected randomly, the relief algorithm then tries to find the two nearest neighbors called respectively *nearest hit H* and *nearest miss M* (line 4). Then the evaluation of the quality $W[A]$ for each feature A based on the attribute values for the R_i, M and H is updated (lines 5, 6). If the attribute value A of the two different instances, such as R_i and H are dissimilar, then the attribute A carry out the separation of the two different instances with the undesirable identical class, such that the quality estimation $W[A]$ is minimized. On the other hand, if the attribute value A of the two different instances, such as R_i and M are dissimilar, then the attribute A carry out the separation of the two instances with desirable dissimilar class values, hence the quality estimation $W[A]$ is reduced. The complete process is reiterated for *m* times where the value m signifies the user-defined factor [22]. Algorithm 1 illustrates the basic Relief algorithm.

ALGORITHM 1. PSEUDO CODE OF RELIEF ALGORITHM

Pseudo code of Relief algorithm
Input: **For all training instance a vector of attribute values and class values**
Output**: vector W of assessment of the attribute qualities**
Assign weights $W[A] := 0.0$;
F**ori:=1 to *m do begin***
randomly choose an instance Ri
Compute closest hit H and miss M
F**orA:=1 to *a* do**
$$W[A] := W[A] - \frac{diff(A,\ R_i, H)}{m} + \frac{diff(A,\ R_i,\ M)}{m}$$
End

Where the function $diff(A,\ I_1,\ I_2)$ is focused on the calculation of the dissimilarity among the attribute value A for two (2) instances I_1 and I_2. For the nominal attribute values, it is presented below as follows:

$$diff(A,\ I_1,\ I_2) = \begin{cases} 0; & Value(A,I_1)=value(A,I_2) \\ 1; & Otherwise \end{cases} \quad (2)$$

And for the numerical attributes is presented as follows:

$$diff(A,\ I_1,\ I_2) = \frac{|value(A,I_1)-value(A,I_2)|}{max(A)-min\ (A)} \quad (3)$$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

To find the closest neighbors the diff function is used for computing the distance among the instances. The total distances over every attribute represents the total distance (Manhattan distance) [22]. However, Relief algorithm has some limitations; in fact, it is not able to handle incomplete data and is restricted for the two-class problems. ReliefF algorithm [22] is utilized for resolving the issues.

ReliefF extended:

The ReliefF algorithm is an advanced version of relief algorithm. It has an advantage to be more robust and capable to manage incomplete and noisy data. Similarly, to Relief, ReliefF randomly selects an occurrenceR_i (line 3), but then searches for k of its nearest neighbors from the same class, called nearest hits H_j (line 4), and also k nearest neighbors from each of the different classes, called nearest misses $R_i(C)$ (Line 5 and 6). It updates the quality estimation $W[A]$ for all attributes A depending on their values forR_i, H_j[22]. Algorithm 2 illustrates the Pseudo code of ReliefF algorithm.

ALGORITHM 2. PSEUDO CODE OF RELIEFF ALGORITHM

Pseudo code of ReliefF algorithm
Input: for all training instance a vector of attribute values and the class value
Output: the vector W of evaluation of the attribute quality
Compute weights $\boldsymbol{W[A] := 0.0}$ **;**
For $\boldsymbol{i := 1}$ **to** m **do begin;**
Choose an instance $\boldsymbol{R_i}$ randomly
Compute $\boldsymbol{k}$ nearest hits $\boldsymbol{H_j}$;
For all class $\boldsymbol{C \neq class(R_i)}$ **do;**
From class $\boldsymbol{C}$ **find** $\boldsymbol{k}$ **nearest misses** $\boldsymbol{M_j(C)}$;
For $\boldsymbol{A := 1}$ **to** a **do**
$$\boldsymbol{W[A] := W[A] - \sum_{j=1}^{k} \frac{diff(A, R_i, H_j)}{m \cdot k} +}$$
Compute $\sum_{C \neq class(R_i)} [\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff\left(A, R_i, M_j(C)\right)] / (m \cdot k)$;
End

Fisher Score (F-score)

F-score is a simplified filter method which computes the separation of two different groups of real numbers [17]. Given training vectors $X_K, K = 1,2,3, \ldots m$, if the overall positive and negative instance is n_+ and n_-, then the F-score of i^{th} feature is expressed as follows:

$$F(i) \equiv \frac{(\overline{X}_i^{(+)} - \overline{X}_i)^2 + (\overline{X}^{(-)} - \overline{X}_i)^2}{\frac{1}{n_+ - 1}\sum_{k,i}^{n_+}(X_{k,i}^{(+)} - \overline{X}^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_- - 1}(X_{k,i}^{(-)} - \overline{X}_i^{(-)})^2} \quad (4)$$

Where $\overline{X}_i, \overline{X}_i^{(+)}$ and $\overline{X}^{(-)}$ denote the i^{th} feature's averages of entire, positive and negative data sets: $X_{k,i}^{(+)}$ denote i^{th} feature of K^{th} negative instance. The numerator signifies the discrimination among positive and negative sets, and the denominator signifies the one within each two sets. The higher the F-score is, the more possible this feature is highly discriminative [17].

2. Wrapper method:

In the wrapper technique, the induction algorithm is utilized for performing selection of features subsets. The induction algorithm is regarded as a black box since knowledge of the algorithm is not required in the beginning of the procedure. Generally, the results provided by wrapper methods are more accurate than filter methods [24]. Figure 2 illustrates the Wrapper approach for feature selection.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

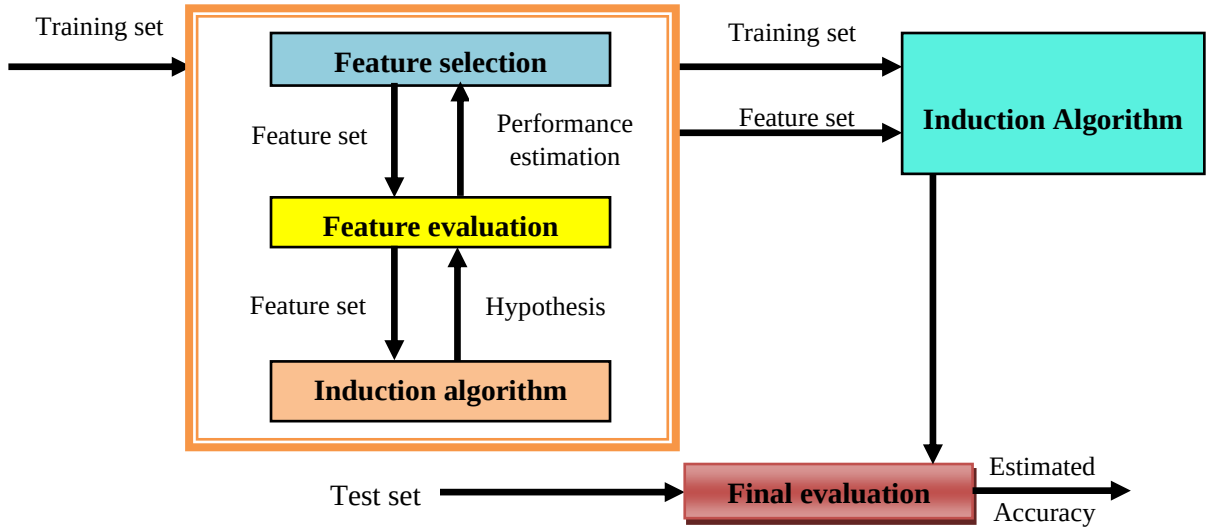


Fig. 2 Wrapper technique for feature selection

C. *Classification with SVMs*

SVM for binary classification was developed in [29]. Considering the training vectors $\boldsymbol{x_i} \in \Re^n, 1 = 1 \ldots m$and a vector of labels $\boldsymbol{y}\epsilon\Re^m, y_i\epsilon\{-1,+1\}$, the optimal hyperplane$f(x) = \boldsymbol{w}^T \cdot \boldsymbol{x} + b$ aiming for separating the training patterns is provided by SVM. Regarding classes that are linearly separable, the summation of distances with respect to the neighboring positive and negative training patterns also named margin is maximized by this hyperplane. To be constructed, the maximum margin needs a correct classification of the vectors x_i of the training set into two dissimilar classes y_i that used the minimum coefficient norms **w** [19].
SVM carry out a mapping of data points within the higher dimensional space $\mathcal{H}$, where the partitioning hyperplane with higher margin is built, when it comes to a non-linear classifier. T quadratic optimization question is given below as,

$$Min_{w,b,\xi}\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\xi_i, \tag{5}$$

Subject to

$$y_i \cdot (W^T \cdot \phi(X_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \tag{6}$$

$\xi_i \geq 0, \qquad i = 1, \ldots m,$

where, the function $\boldsymbol{X} \rightarrow \phi(\boldsymbol{X})\epsilon\mathcal{H}$ maps training data to the higher dimensional space$\mathcal{H}$. An ensemble of loose variables ξ employed for every training vector and C signifies the penalty parameter on training the error [29] [19].

The SVM mapping has the following solution:

$$f(\boldsymbol{X}) = sign(\sum_{i=1}^{m} y_i\alpha_i^*(\boldsymbol{X}) \cdot \phi(\boldsymbol{X_i}) + b^*) \tag{7}$$

The mapping function $K(\boldsymbol{X},\boldsymbol{Y}) = \phi(\boldsymbol{X}) \cdot \phi(\boldsymbol{Y})$ is implicitly specified by the kernel function K utilized for computing the inner product of two vectors in$\mathcal{H}$. Thus, decision function is illustrated as follows [19]:

$$f(\boldsymbol{X}) = sign(\sum_{i=1}^{m} y_i\alpha_i^*K(\boldsymbol{X},\boldsymbol{Y}) + b^*) \tag{8}$$

The maximum margin is the maximal distance in the feature space $\mathcal{H}$ to the nearest image $\phi(\boldsymbol{X_i})$ from the training data and the double formulation can be specified in the following manner [19]:

$$Max_\alpha \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,s=1}^{m}\alpha_i\alpha_s y_i y_s K(\boldsymbol{X}_i,\boldsymbol{X}_s), \tag{9}$$

Subject to
$\sum_{i=1}^{m}\alpha_i y_i = 0,$
$0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, m.$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

Moreover, the three various kinds of commonly utilized kernel functions [23] are:

- Linear kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_s) = \boldsymbol{X}_i \cdot \boldsymbol{X}_s$ (10)
- Polynomial kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_s) = (\boldsymbol{X}_i \cdot \boldsymbol{X}_{s+1})^d$ (11)
- Gaussian kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_s) = \exp\left(-\|\boldsymbol{X}_i - \boldsymbol{X}_s\|^2 / 2\sigma^2\right)$ (12)

where the order *d* of the polynomial kernel equation (10) and σ of Gaussian kernel equation (11) are the adaptable kernel parameters.

D. Cross validation

Cross-Validation (CV) is extensively utilized for estimating the performance of various applications in general and specifically with SVM and Gradient Boosted trees. The major objective of CV is to partition the data, once or several times, to estimate the threat inherent to each algorithm. The CV uses part of data (the training sample) to train every algorithm, and also uses the left-over component (the validation sample) to estimate the risk of the algorithm. At last, the algorithm with the smallest estimated risk is selected by CV [1].

K-fold CV is considered as the basic form of CV. The procedure behind a classic K-fold CV for the linear model is characterized by random and even split of data set into K parts (If possible). The procedure behind a classic K-fold CV for the linear model is characterize by random and even split of data set into K parts (If possible). K-fold CV builds a candidate model with respect to the K-1 parts of data set, named training set. The evaluation of the prediction accuracy of candidate model is performed on the test set that includes the data in hold-out division. The optimal model considered as the model having minimum CV score (usually, the mean-squared prediction error (MSPE)) is selected using all the K parts as the test set and reiterating the model construction and assessment procedure. Let be *p* independent variables, in all there are $2^p - 1$ possible models. Every model is evaluated K times in k-fold CV practice. Thus, the selection of a unique 'optimal' model is performed through K(2^p-1) times of the model assessment [12].

Generally, the data are portioned into *k* equally sizes segments or folds. Hence 10-fold cross validation (k=10) is utilized in the research because of fact that it is commonly applied for machine learning and data mining techniques and it is applied on SVM and Gradient Boosted Trees for the feature classification.

IV. RESULTS AND DISCUSSION

The results and discussion of developed NIDS approach based on the SVM and Gradient Boosted Trees using the performance metrics is illustrated in this subsection.

A. Experimental Setup

The experimental implementation of proposed NIDS approach based on the SVM and Gradient Boosted Trees is performed using the NSL-KDD data set. Experiments are conducted with different weights for both Fisher score ranks and ReliefF score ranks. Calculations are conducted on selected features using SVM and Gradient Boosted Trees.

B. Dataset description

The NSL-KDD [7] is a latest version of the dataset KDD which is used for solving various inherent issues of KDD'99. The advantages reported by the NSL KDD dataset are:

- First it does not include unnecessary reports in the training sets, such that the classifier does not generate any partial results.
- Second reports are not duplicated in the testing set, and hence achieved efficient reduction rates.
- Third the total number of the chosen records from every complex level group is inversely proportional to the record percentage in the real dataset KDD

The NSL-KDD dataset is made up of 41 features and 5 normal classes and 4 kinds of attacks: User to Root (U2R), Remote to Local (R2L), Probe and Denial of Service (Dos).

U2R: U2R is a kind of attack where an attacker utilizes a legitimate account for entering into the victim system and intends to gain privileges as an administrator (root). To do this the attacker exploits some vulnerability in the victim's post.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

R2L: Remote to Local is a kind of attack where the intruder access into the remote machine and enhances the local admittance of the victim machine.
Probing: Probing is an attack category that aims to gain the victim information remotely.
DOS: Denial of Service is the attack where the victim's resources are exhausted, thus building the system incapable of handling authorized requests.

1. Analysis based on features

Features protocol type, service, flag get strings values, and features protocol type gets values: tcp, udp, icmp. These features are recorded in the following way: feature protocol type has spread to three different features, such as protocol type udp, protocol type tcp, protocol type icmp, and these three new features get just [0,1] values.

If protocol type is equal to tcp and feature protocol type tcp is equal to 1, then protocol type udp get 0, and protocol type icmp get 0 as well. This process is called dummy coding. The feature protocol type is erased after dummy coding and just the three new features stay, namely protocol type udp, protocol type tcp, protocol type icmp. This type of features is called categorical features and the dataset contains three different categorical features, such as protocol type, flag, and service.

Categorical features have been dummying coded and erased afterwards. Class feature values are changed from [normal, anomaly] to [0,1]. All features are scaled since SVM requires scaling, and the standard scaling are used for both train and test split. The standard scaling is a process where from every row's item is subtracted column mean and divided by column standard deviation. The column mean means train set column mean and the column standard deviation means train set standard deviation. These values are utilized for the test set scaling. The standard scaled data (columns) have mean value equals to 0 and standard deviation equals to 1.

$$\frac{x-\bar{x}}{s} \tag{13}$$

Analysis based on features with respect to Fisher score and SVM:

Sorted lists of features are set according to their Fisher score in descending order. Five folds cross fold validation are run for all the models. Here, the first model is the one with a single variable with the maximum Fisher score. The second model is the one with the first two features with the maximum Fisher Score, and third model is the model with three features with maximum Fisher score and final model contains all the features. Fisher score described in [25] is used, whereas the SVM with RBF kernel with default settings includes no hyper parameter tuning. In addition, accuracy [7] is considered as a measure of quality of model.

C. Performance metrics

This section describes the evaluation metric, namely accuracy for performing the method analysis.

1. Accuracy for fold=accuracy (mean (fold)):

The focus is on algorithms for representing the flat features, where the features are assumed to be independent, specifically in filter models. Filter models determine the features without using any categorization algorithms and rank features on certain criteria. The novelty in this approach is that weighted rank is calculated on all the features. Fisher score filter model and reliefF filter model are used, ranks are assigned to features. Afterward, weighted ranks are calculated for all the features.

$$Weighted\ Ran\ k_i = w_a.FisherScoreRank_i + w_b.reliefFScoreRank_i \tag{14}$$

where,

$$[w_a, w_b] = \left[\frac{1}{3}, \frac{2}{3}\right], \left[\frac{2}{3}, \frac{1}{3}\right], \left[\frac{1}{4}, \frac{3}{4}\right], \left[\frac{3}{4}, \frac{1}{4}\right] \tag{15}$$

$$[w_a, w_b] = \ combineda1, combineda2, combineda3, combineda4 \tag{16}$$

Features are ordered in descending order according to their weighted ranks and Support Vector Classifier (SVC) models is calculated. $1 \ldots m$. Features used in SVM models and the accuracy are calculated from holdout set predictions and test set predictions.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

D. Comparative Analysis

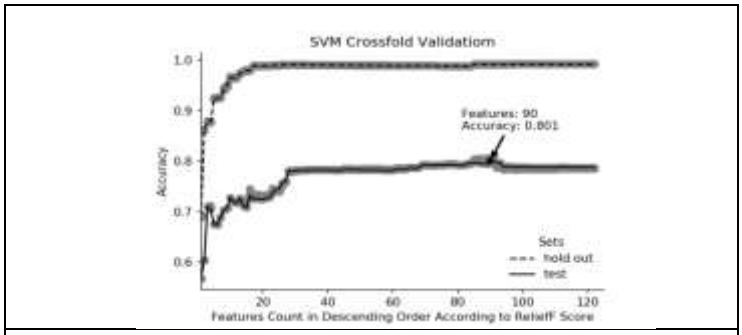
The comparative analysis of the developed NIDS approach based on the SVM and Gradient Boosted Trees based on the evaluation metric, such as the accuracy by considering the NSL-KDD database is explained in this section.

1. Analysis based on SVM calculation plots:

This section illustrates the analysis of the SVM technique with respect to ReliefF score, Fisher score, and combined scores based on features with respect to accuracy.

Analysis using ReliefF score:

Figure 3 portrays the analysis based on ReliefF score with respect to accuracy. Figure 3a) illustrates the accuracy assessment with respect to varying feature counts. By considering the feature as 90, the accuracy value obtained by the SVM cross fold validation is 0.801. The SVM cross fold validation achieved an accuracy value of 0.780 for the feature count 70.

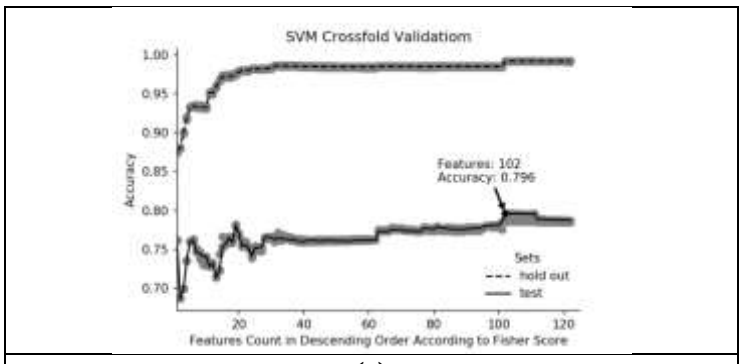


(a)

Fig. 3 Analysis based on features according to reliefF score a) Accuracy

Analysis based on Fisher score:

The analysis based on Fisher score using the accuracy metric is presented in figure 4. Figure 4a) portrays the assessment of accuracy metric. For the feature count 102, the accuracy value measured by the SVM cross fold validation is 0.796. The SVM cross fold validation achieved an accuracy of 0.775 for the feature count 20.



(a)

Fig. 4 Analysis based on features according to Fisher score a) Accuracy

Analysis based on combined ReliefF score and Fisher score:

This section illustrates the combined score values with respect to ReliefF score and Fisher score based on the feature count with respect to accuracy. Figure 5a) illustrates the analysis based on ranked combined a1 score. For the feature count 71, the accuracy value achieved by the SVM cross fold validation is 0.803. The analysis with respect to the ranked combined a2 score is portrayed in figure 5b). When the feature count is 84, the SVM cross fold validation measured an

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

accuracy value of 0.797. Figure 5c) depicts the assessment based on ranked combined a3 score. The SVM cross fold validation obtained an accuracy value of 0.803 for the feature count 77. The analysis according to the ranked combined a4 score is portrayed in figure 5d). When the feature count is 103, the SVM cross fold validation obtained an accuracy of 0.796 respectively.

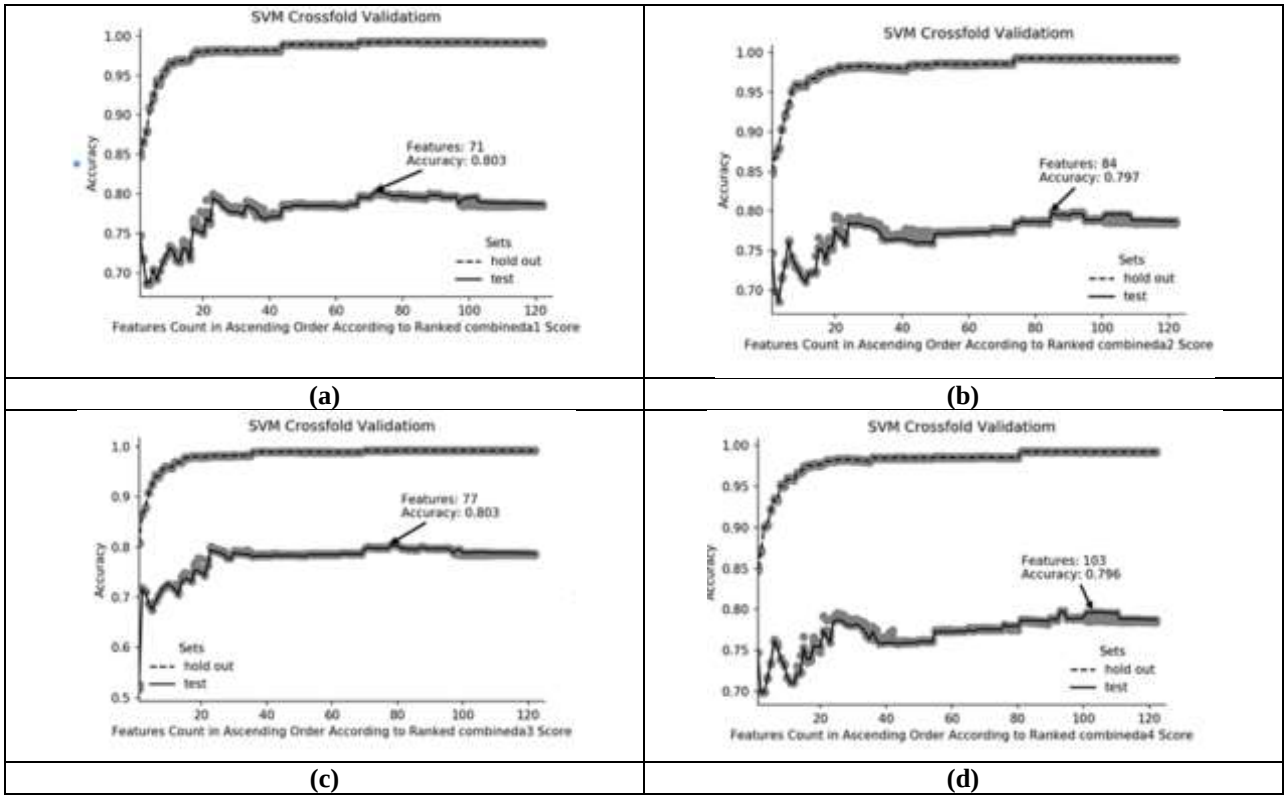


Fig. 5 5a), 5b), 5c) and 5d) show features according to combined methods

Analysis based on Gradient Boosted Trees:

This section illustrates the analysis of the Gradient boosted tree with respect to features according to the wrapper method.

Analysis using wrapper technique:

Figure 6 portrays the analysis of the feature count in decreasing order with respect to Wrapper technique. Figure 6a) portrays the assessment of the Gradient Boosted Tree Crossfold validation based on accuracy metric. For the feature count 37, the Gradient Boosted Tree Cross fold validation measured an accuracy of 0.567. When the feature count is 47, the accuracy value obtained by the Gradient Boosted Tree Cross fold validation is 0.551.

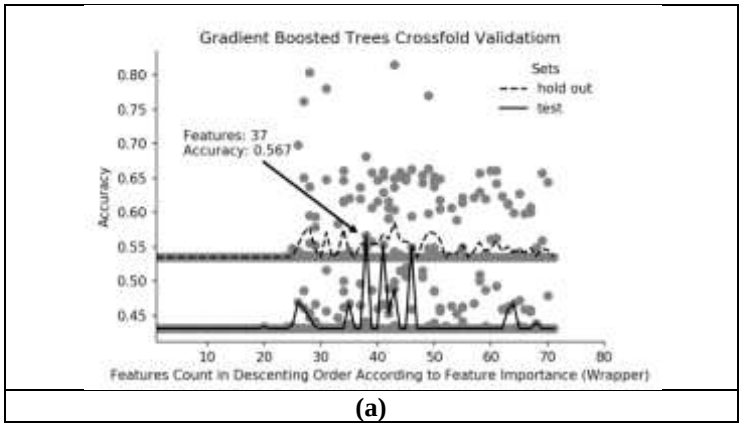


Fig. 6 Features importance for Gradient Boosted Trees according with Wrapper method

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

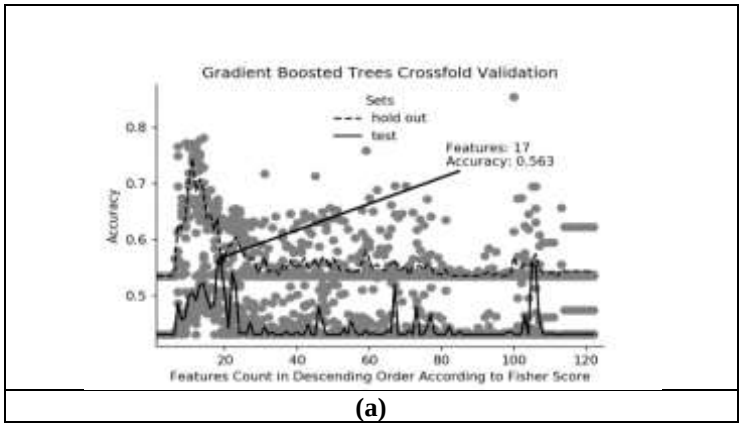
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

Analysis based on Fisher score with respect to 17 features:

Figure 3 portrays the analysis with respect to Fisher score using 17 features based on accuracy metric. Figure 3a) illustrates the accuracy assessment with respect to varying feature counts in decreasing order. By considering the feature as 17, the accuracy value obtained by the Gradient Boosted Tree Crossfold validation is 0.563. The Gradient Boosted Tree Crossfold validation achieved an accuracy value of 0.521 for the feature count 67.

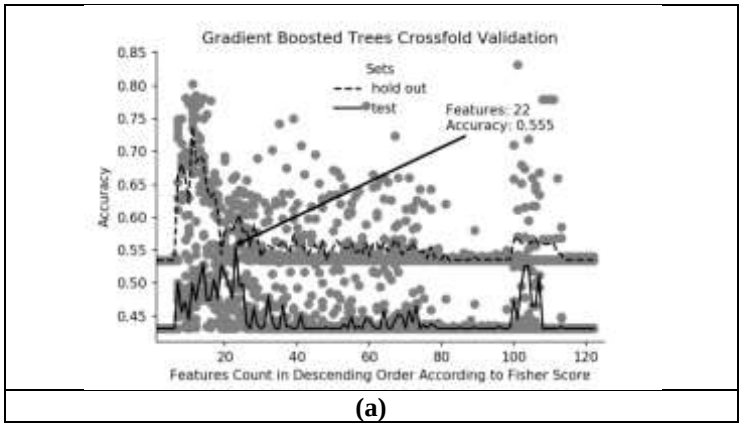


(a)

Fig. 7 Features importance for Gradient Boosted Trees according to Fisher Score: 17 features.

Analysis based on Fisher score with respect to 22 features:

The analysis based on Fisher score using the accuracy metric is presented in figure 4. Figure 4a) portrays the assessment of accuracy metric. For the feature count 22, the accuracy value measured by the Gradient Boosted Tree Crossfold validation is 0.555. The Gradient Boosted Tree Crossfold validation achieved an accuracy of 0.522 for the feature count 102.



(a)

Fig. 8 Features importance for Gradient Boosted Trees according to Fisher Score: 22 features.

V. CONCLUSION

This paper develops a robust model for the intrusion detection systems using the SVM and Gradient Boosted Trees. Here, features selection is performed by different approaches, with first filter method using Fisher score, second wrapper method and third the combination of Fisher score and ReliefF score. The combination of filters improved the results in a way that a smaller number of features is needed to achieve good results. This method achieved higher accuracy during intrusion detection in a network-based computer system while using filter and wrapper techniques in one hand and the integration of fisher score and ReliefF score in the other hand. Hence the method has advantages over other approaches. The model is successfully tested with NSL-DKK dataset. The limitation of this method is that it becomes computationally extensive as complexity is growing and implementation might be difficult in a production system. The future work would be the concern of improving the approach by the linear combination of Fisher score rank, ReliefF score rank, like wrapper rank or other known approach.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10621

REFERENCES

[1] Arlot, S., & Celisse, A., "A survey of cross-validation procedures for model selection", Statistics Surveys, vol.4, pp.40-79, 2010.
[2] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K., "Network anomaly detection: Methods, systems and tools", IEEE Communications Surveys and Tutorials, vol.16, no.1, pp. 303–336, 2014.
[3] Chandola, V., Banerjee, A., & Kumar, V., "Anomaly Detection: A Survey", ACM Computing Surveys, vol.41, no.3, pp.1–58, 2009.
[4] Chandrashekar, G., & Sahin, F., "A survey on feature selection methods", Computers and Electrical Engineering, vol.40, no.1, pp.16–28, 2014.
[5] Claeskens, G., & Kerckhoven, J. Van., "An Information Criterion for Variable Selection in Support Vector Machines", Journal of Machine Learning Research, vol.9, pp.541–558, 2008.
[6] Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K., "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks", Expert Systems with Applications, vol.29, no.4, pp.713–722, 2005.
[7] Dhanabal, L., & Shantharajah, S. P., "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, vol.4, no.6, pp.446–452, 2015.
[8] Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., & Petrovski, K. R., "Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models", Computers in Biology and Medicine, vol.114, pp.103456, 2019.
[9] Hamed, T., Dara, R., & Kremer, S. C., "Network intrusion detection system based on recursive feature addition and bigram technique", Computers and Security, vol.73, pp.137–155, 2018.
[10] Hu, Z., Bao, Y., Xiong, T., & Chiong, R., "Hybrid filter-wrapper feature selection for short-term load forecasting", Engineering Applications of Artificial Intelligence, vol.40, pp.17–27, 2015.
[11] Huang, J. J., & Chen, C. Y., "Integration of rough sets and support vector machines for network intrusion detection", Journal of Industrial and Production Engineering, vol.31, no.7, pp.425–432, 2014.
[12] Jung, Y., & Hu, JA., "K-fold averaging cross-validation procedure", Journal of Nonparametric Statistics, vol.27, no.2, pp.167–179, 2015.
[13] Khammassi, C., & Krichen, S. A., "GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection", Computers & Security, 2017.
[14] Kim, G., Lee, S., & Kim, S., "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", Expert Systems with Applications, vol.41, no.4, pp.1690–1700, 2014.
[15] Kira, K., & Rendell, L. A., "The Feature selection Problem: Traditional Methods and New Algorithm", AAAI-92 Proceedings, 1992.
[16] Kohavi R, John GH., "Wrappers for feature subset selection", Artificial intelligence, vol.97, no.1-2, pp.273-324, December 1997.
[17] Lee, M. C., "Using support vector machine with a hybrid feature selection method to the stock trend prediction", Expert Systems with Applications, vol.36, no.8, pp.10896–10904, 2009.
[18] Liu, Y., & Zheng, Y. F., "FS _ SFS : A novel feature selection method for support vector machines", Pattern Recognition, vol.39, pp.1333–1345, 2006.
[19] Maldonado, S., & Weber, R., "A wrapper method for feature selection using Support Vector Machines", Information Sciences, vol.179, no.13, pp.2208–2217, 2009.
[20] Maldonado, S., Weber, R., & Basak, J., "Simultaneous feature selection and classification using kernel-penalized support vector machines", Information Sciences, vol.181, pp.115–128, 2011.
[21] Ring, M., & Eskofier, B. M., "An approximation of the Gaussian RBF kernel for efficient classification with SVMs", Pattern Recognition Letters, 2016.
[22] Robnik-Sikonja, M., & Kononenko, I., "Theoretical and Empirical Analysis of ReliefF and RReliefF", Machine Learning, vol.53, pp.23–69, 2003.
[23] Shieh, M. D., & Yang, C. C., "Multiclass SVM-RFE for product form feature selection", Expert Systems with Applications, vol.35, no.1–2, pp.531–541, 2008.
[24] Sun, Y., "Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications", IEEE transactions on pattern analysis and machine intelligence, vol.29, no.6, pp.1035–1051, 2007.
[25] Tang, J., Alelyani, S., & Liu, H., "Feature selection for classification: A review", In Data Classification: Algorithms and Applications, 2014.
[26] Tu, I. B., "A novel feature selection approach: Combining feature wrappers and filters", Information Sciences, vol.177, pp.449–466, 2007.
[27] Unler, A., Murat, A., & Babu, R., "Mr^2 PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification", Information Sciences, vol.181, no.20, pp.4625–4641, 2011.
[28] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H., "Relief-based feature selection: Introduction and review", Journal of Biomedical Informatics, vol.85, pp.189–203, 2018.
[29] Vapnik, V. N., "Statistical Learning Theory", 1998.
[30] Varshavsky, R., Gottlieb, A., Linial, M., & Horn, D., "Novel Unsupervised Feature Filtering of Biological Data", vol.22, no.14, pp.e507–e513, 2006.
[31] Wang, G., Chen, S., & Liu, J., "Anomaly-based Intrusion Detection using Multiclass-SVM with Parameters Optimized by PSO", International Journal of Security and its Applications, vol.9, no.6, pp.227–242, 2015.
[32] Wang, M., & Chen, H., "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis, Applied Soft Computing Journal, vol.88, pp.105946, 2019.
[33] Zavrak, S., & Iskefiyeli, M., "Anomaly-Based Intrusion Detection from Network Flow Features Using Variational Autoencoder", IEEE Access, vol.8, pp.108346–108358, 2020.