Building attack detection system base on machine learning

Rasha Thamer Shawe*, Kawther Thabt Saleh and Farah Neamah Abbas

Department of Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq

Global Journal of Engineering and Technology Advances, 2021, 06(02), 018–032

Publication history: Received on 09 January 2021; revised on 06 February 2021; accepted on 08 February 2021

Abstract

These days, security threats detection, generally discussed to as intrusion, has befitted actual significant and serious problem in network, information and data security. Thus, an intrusion detection system (IDS) has befitted actual important element in computer or network security. Avoidance of such intrusions wholly bases on detection ability of Intrusion Detection System (IDS) which productions necessary job in network security such it identifies different kinds of attacks in network. Moreover, the data mining has been playing an important job in the different disciplines of technologies and sciences. For computer security, data mining are presented for serving intrusion detection System (IDS) to detect intruders accurately. One of the vital techniques of data mining is characteristic, so we suggest Intrusion Detection System utilizing data mining approach: SVM (Support Vector Machine). In suggest system, the classification will be through by employing SVM and realization concerning the suggested system efficiency will be accomplish by executing a number of experiments employing KDD Cup’99 dataset. SVM (Support Vector Machine) is one of the best distinguished classification techniques in the data mining region. KDD Cup’99 data set is utilized to execute several investigates in our suggested system. The experimental results illustration that we can decrease wide time is taken to construct SVM model by accomplishment suitable data set pre-processing. False Positive Rate (FPR) is decrease and Attack detection rate of SVM is increased .applied with classification algorithm gives the accuracy highest result. Implementation Environment Intrusion detection system is implemented using Mat lab 2015 programming language, and the examinations have been implemented in the environment of Windows-7 operating system mat lab R2015a, the processor: Core i7- Duo CPU 2670, 2.5 GHz, and (8GB) RAM.

Keywords: Attack Detection0, Intrusion Detection System (IDS)0 ,Data Mining0, Support Vector Machine (SVM)0.

1. Introduction

With the rise utilize the networkeds computerrs fors crucial systemss ands thes common utilize of distributed ands large computers networks, the security of computers networks concern rises and network intrusions have been a dangerous risk in latest time. Intrusions detections systems (IDS) hass beenes great used to be a seconds row of protection form networked computers systems sidways with additional network security methods for instance access controland firewall. The majoraimof IDS is to detect illegal utilize,abuse and misuse of computers systems by boths systems insiderrands outsiders intruders. Theres are differentmethodssto construct intrusions detections systems (IDS). IDSs can bes classifieds into twos classifications depend on the approachess utilized to detects intrusions: abuse detectionsand anomalysdetection[1, 2, 3]. Anomalysdetectionsmethodscreatessthe profilesof sosusual actions ofs users, system resources, operatings systems, networks services ands traffic usingthe examination trails created by anetwork scanning program or a host operating system. This method detects intrusions by classifying important perversions from the usual attitude samples of these profiles. Anomaly detection method is not necessary that previous knowing of the security holes of the goal systems. So, this system is capable tosdetect not onlyidentified intrusionssbut also unidentiefiedintrusions. Moreover,thissmethods can identify the intrusionssthatsare accomplished by the misuse

* Corresponding author: Rasha Thamer; Email: rashatham82@uomustansiriyah.edu.iq
College of Education in Computer Science AL-Mustansiriya University Baghdad, Iraq.

Copyright © 2021 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0.
of legal users or disguise without violation security politics [4,5,26]. The drawbacks of this method were it had rise fakespositive recognition fault, the hardness of amenable progressive misbehavior, and costly calculation [4,6,7]. otherwise, misuse recognition method determines doubtful abuse signatures depended on known system weaknesses and a security procedure. Abuse method achieve whether signatures of identified attackss are existing or notsin the auditing path, and ansy corresponded behavior is recognized as attack. Misuses detection method identifiessonly previously recognized interference signatures. The benefit of this method is that sit seldom defeat to identify prior told intrusions, s is decrease fake positives rate [5,8]. Thedifficulties of this method cannot identify modern intrusions it have not ever before been detected, i.e. Greater fake negatives rate. Moreover, this method hassanother disadvantages as the hardness of misuse signature bases and the hardness of creating and updating signature rules of intrusion [4, 8, 9]. These are twotypes of Intrusion detection systems IDS are Networks Intrusion Detection System NIDS and Host-based Intrusion Detection (HIDS) [9,10,27]. In our study we built Intrusion detection systems (IDS) based on data mining.

Intrusion detection systems (IDS) approachs are two complementary orientations in intrusion detection [11,12,28]: Misuses detection. These seeking for proof of attacks depend on the information collecteds from recognized attacks. Moreover, it is indicate to attackss type as detections by appearance or misuse detection. Anomaly detection is seeking for perversion from the pattern of uncommon attitude depend on the monitoring of a system during a ordinary status and is indicated to such anomaly detection or find via behavior.

2. Related work

2.1. Soni and Sharma in 2014[13]

suggested two techniques artificial neural network (ANN) and C5.0 are employed together with characteristic picking. Feature picking method eliminate several irrelevant characteristics while C5.0sand ANN performed such a classifiersto categorizes the input data in eithernormal category or attack that ones of the fivetypes. KDD99sdataset is employed to test and train the system; C5.0ssystem through numbersof characteristics is make improved results with nearly 100% accuracy. Moreover, they used ANN approach to categorize intrusion data depend on their partition size. A comparative result demonstrates that C5.0 is execution better than ANN and yields best outcome with 36 features.

2.2. Zargar and Baghaie in 2012[14]

offered a category-basedpicking of active parameters for detection of intrusion utilizing Principals Components Analysis (PCA). They employ 32main characteristics from Transmission Control Protocol// Internet Protocol (TCP/IP) header, also 116resulted features from TCPdump are picked in adataset of networktraffic. Attackss are classified in fivetypes, User attack (U2R), Denialsof Services (DoS), Remotesto and Probingattack, Remote Usersattack (R2L). Moreover, they used TCP dump froms DARPA 1998dataset insteas the tests as the picked dataset. PCA approach is utilized to define an ideal characteristic setsto produce the detection procedure higher speed. The experimental results display that characteristic reductions can get better detections rate for the category-based detection method while the continuing detection accuracy within a suitable range. The KNNs classification technique is utilized for the attackss classification. The experimental results illustrate that characteristic reductions will importantly speed up the testing and training the time for recognition of the intrusion challenges.

2.3. Mukkamala and Sung in 2003[15]

proposed Feature picking for Intrusion Detection utilizing Two learning machine classes for intrusion detection system (IDS) are studied: Artificial Neural Networks (ANNs) and Support Vectors Machines (SVM). They display that SVMs are better than ANNs three serious respectss of intrusion detection system: SVM execute and train are greatness quicker; SVMs scale much superior; and SVM provide greater classification accuracy. Moreover, address the concerning matter of ranking the significances of input characteristics, which is a hazared of major significant. Sincereomoval of the useless and/or unsimonials inputss produce a simplified hazared and probably quicker and extra precise detection, characteristic picking is quite significant in intrusion detection. The experimental results show that SVM-depend IDS utilizing a reduced feature numbers can deliver improved or comparable performance. In conclusion, IDS suggested depend on SVM for detecting an exact category.

2.4. Zhu et al.,2005 [16]

RICGAs (ReliefF-Immunes Clonal Genetics Algorithm), a collective characteristic subsets picking method depend on the Immune Clonal selection, ReliefF algorithm and GA is suggested in is employed BP networks as classifier. RICGA has
higher accuracy of classification (86.47%) for small scope characteristic subsets than ReliefF-GA. In the paper, the features are not mentioned. A composite characteristic subset selection method, claimed RICGAs (ReliefFimmune clonal genetic algorithm), depend on the immune clonal selection algorithm, ReliefF algorithm and GA. In the RICGA method, they employed in the first ReliefFsto get rid of irrelevant characteristics, then execute a improved genetic algorithm to get the lastly characteristic subset. They analyse hardly the RICGA Markov chain model algorithm and its convergence. Experimental results on real KDD CUP’99 datasets display that the RICGAsmethodsis eminent to the ReliefF-GAs and GAs on classifications precision and input characteristic subset size.

2.5. Ming-Yang Su (2011) [17]

Offered an approach for featurespicking to identify DoS/DDoSattacksfor designing in realtime anomaly-based Network Intrusion Detection System NIDS. Genetic algorithms (GA) collective with KNN (k-nearest-neighbor) are utilized for characteristic weighting and picking. The outcome of KNN’s classification is employed as the fitness function in a GA to improve the characteristics weight vectors. First 35 characteristics in the train stage are weighted. The highest 19 characteristics are taking into account for recognized attacks and the top 28 characteristics for unidentified attacks. In this paper, extracted characteristics are not aforesaid. A total accuracy rate of 97.42% is obtained for recognized attacks and 78% for unidentified attacks.

3. Data Mining and Intrusion Detection Systems

Intrusion detection systems (IDS) have been depend intraditionally on the feature of an attack and the system tracking activity to check if it matches that description. IDS depend on data mining is creation their appearance more ability. The system of Data Mining approaches for intrusion detection applications have been commonly employed these days. The intrusion detection difficult has been reduced to a Data Mining mission of classifying data. Summarized, given a data points set belonging to various attacks activity (9 classes) and purposes to isolate them as accurately as possible by means of a model. Many various data mining approaches have been found for intrusion detection classification. In this system, we employed a Support Vector Machine (SVM) for attack detection as a classification algorithm. Also, we used feature extraction and dimensionality reduction algorithms (PCA and LDA, SVD) based on the KDD’99 Cup datasets.

4. Design and Implementation of Proposed System

The proposed intrusion detection system to scheme a proficient intrusion detection and recognition system is described as follows:

![Figure 10 Intrusion Detection Classification proposed System Approach](image)

The aim of analyses is to increase the intrusion detection system achievement; the data which used as input to proposed system is KDD Cup 99 dataset. The KDD Cup 99 dataset is requirement to pre-processing which is done by converting all data into similar format. Then feature reduction is performed to extract and reduction features. Finally, intrusion classification stage is done by based on different kind of system insertions, the classification algorithms Support Vector
Machine (SVM). As KDD Cup 99 dataset holds some symbolic attribute and also numeric attributes, two sorts of transformation technique have been utilized for these properties. The two machine learning procedures are prepared on both kind of transformed dataset and afterward their outcomes are looked at with respect to the correctness of intrusion detection. The suggested system is containing fundamentally of two essential jobs which are feature reductions and attack detection.

Our proposed intrusion detection system steps are showed in Figure (1), which includes the main parts. Input KDD’99 Dataset, Dataset Pre-processing, Dimensionality Reduction and feature selection, Classification Algorithms and Performance Measurement.

4.1. KDD’99 Input Dataset

In first phase of the suggested intrusion detection system gets the KDD Cup 99 dataset as an input where the whole record numbers. In our proposed system, we utilized the total KDD Cup 99 dataset. Each record on 42 features; the records have labeled either attack or normal type.

The KDDsCUP 1999 [18] standard datasets are employed to evaluate various characteristic selection technique for Intrusion detection system. This system contains of 4,940,000 related records. Every relation had a label of either attack or the normal kind, with quite one exact attack category happens in one of the four attacks types [19] as: User to Root Attacks (U2R), Remote to Local Attacks (R2L), Denial of Service Attack (DoS) and Probing Attack.

Denial of Service Attack (DOS): Attacks of this category deprive the legitimate or host user from utilizing the resources or service.

Probe Attack: Theses attacks mechanically scan a computer networks or a DNS server to get legal IP addresses.

Remote to Local (R2L) Attack: In this attack category an attacker who does not have an account on a victim machine achieves local access to the mechanism and changes the data.

User to Root (U2R) Attack: In this attack category a local user on a mechanism is able to get excellence normally kept for the supers (root) users. Each related record consists of 41 features and also are labeled in configuration ass 1, 2, 3, 4, 5, 6, 7, 8, 9, ..., 41 and fall into the four types are displayed in Table 1:

Category 1 (1-9): Elementary features of separate TCP associates.

Category 2 (10-22): Implicate features within an association proposition by domain knowledge.

Category 3 (23-31): Traffic features calculated utilizing a two-second time window.

Category 4 (32-41): Traffic features calculated utilizing a two-second time window from goal to host.

Table 1. Distribution of intrusion Types in datasets

| Dataset          | Normal00 | Probe00 | DOS00  | U2R00 | R2L00 | Total00 |
|------------------|----------|---------|--------|-------|-------|---------|
| KDD cup data     | 97280    | 4107    | 391458 | 52    | 1124  | 494020  |

4.2. KDD’99 Pre-processing

KDD’99 pre-processing is second phase and is one of the significant phases of system. This stage proper data to be accepted to next phase for extraction and reduction data. This phase contain from two step (Dataset Labeling, Normalization). The following subsection will illustrate all details about these steps.

4.2.1. Dataset Labeling

The Dataset Labeling is the first step in KDD’99 Pre-processing phase. This step is so important. The output of Dataset Labeling employed as input to next step in Pre-processing phase (Normalization). The dataset labeling is done by utilizing the whole features in the KDD 10% corrected datasets at it displayed in the screen shot which is sited in the
second cell of the entire dataset. The figure (2) is the KDD 99 dataset screenshot that we took it from environment of our matlab.

![Feature No (3)](image1)

![Feature No (2)](image2)

![Feature No (42)](image3)

**Figure 2** First KDD cup dataset row of 10% correction (data sample)

So, the Table (2) illustration labeled the dataset base on attacks which are fall into one of five categories as below:

**Table 02** Our Class labeling of “10% KD099” dataset

| Attackss Types | Descriptions                                                                 | Subs Types           | Labels |
|----------------|------------------------------------------------------------------------------|----------------------|--------|
| (0DoSs0) Denials of sService | Attacker tries to prevents legitimates userss froms using as services | Smurfs               | 1      |
|                |                                                                              | Neptunes             |        |
|                |                                                                              | backs                |        |
|                |                                                                              | Teardrops            |        |
|                |                                                                              | pods                 |        |
|                |                                                                              | lands                |        |
| Normals        | datas with nos attacks                                                      | normals              | 2      |
| Probes         | Attacker tries to prevents legitimates userss froms using as service.        | Satans               | 3      |
|                |                                                                              | ipsweeps             |        |
|                |                                                                              | portsweeps           |        |
|                |                                                                              | nmaps                |        |
| (R2L) Remotes to Locals | Attacker does not have an accounts on the victim machines, hence tries to gains access | warezclients         | 4      |
|                |                                                                              | guess_passwds        |        |
|                |                                                                              | Warezmasters         |        |
|                |                                                                              |Imaps                |        |
|                |                                                                              | ftp_writes           |        |
|                |                                                                              | Multihops            |        |
|                |                                                                              | Phfss                |        |
|                |                                                                              | spys.                |        |
| Users to Roots (U2R) | Attacker has local accesss to the victims machines and tries to gains supers useres privilege | buffer_overflows     | 5      |
|                |                                                                              | Rootkits             |        |
|                |                                                                              | Loadmodules          |        |
|                |                                                                              | Perl                 |        |

The datasets record includes 42 characteristics (e.g. 0_service0, protocoltype0, andsFlag) and is labeled as either attack or a normal also illustrate any one of attack type as presented in Figure (1.2). As an example, if we take a sample of first row from the KDD 99 dataset before doing the normalization. The Figure (2) is clear that the feature numbers is (42) which has the definite attack category as described in Table (2).

The Table (3) illustrate the dataset how must be labeled by employing 10% of the corrected dataset:
There is also another issue in this step. There are many nominal values in the dataset such as HTTP, SF, and ICMP. Consequently in this step transform all nominal values to numeric values in advance. For instance, the service form of "tcp" is mapped to 1, "udp" is mapped to 2, "icmp" is mapped to 3 and the table (3) shows all transformation the dataset nominal value features into the numeric values. Figure (3) has shown the original KDDCUP 1999 dataset will become after transformation as display in figure (3).

**Figure 3** Pre-processing Original KDDCUP 1999 dataset before and after transformation

### 4.2.2. Normalization

After we do the labeled for all dataset feature space, we can do the dataset Normalization by using the whole KDD010% corrected0 datasets at it shown0in the screen shots which is located0in the second cell of the whole datasets.

KDD’99 as an input dataset includes characteristic numbers and these are in0 different style. Some are numbers of style and others are in character style. Consequently, in this stage this various style dataset is transformed into same style to be extracted0to the next phases.
Since there are some KDD CUP 99 dataset features are continuous, therefore a normalize process is done on these features to become more suitable for the DM classification algorithms. Normalization is utilized for preprocessing the data, where the characteristic data are ranges to be in a tiny definite scaled for example 0.00 to 1.00 or -1.00 to 1.00. Normalizing the input values for every characteristic measured in the training patterns will said speeds up the learning phase.

4.3. Features Extraction and Dimensionality Reduction of the KDD990

Features extraction and dimensionality reduction method is done by eliminating redundant and irrelevant features. Irrelevant is that features have little connection with class labels. The redundant features have robust relationship with picked features. In this suggested system we employed three various algorithms which are Singular value decomposition (SVD), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) approaches. We used these techniques for extracting appropriate characteristics from dataset, and reduce dimensions of the KDD as well then are given as an input to a next step.

4.3.1. Principal Component Analysis (PCA)

PCA is a convenient statistical approach that has found systems in fields for instance image compression and face recognition, and is a popular method for definition samples in high dimension data. The whole object of statistics is depend on about the conception which big data set, and examine set describes of the relations between the separate point in that set [20]. The objective of PCA is to limit the dimensionality of the dataset while preserving as far as possible the variance current in the original dataset. It is asway of categorizing samples insdata, and term the data in as a technique as to focus their differences and similarities [21].

Algorithm 1 Principal Component Analysis (PCA)

4.3.2. Singular-Value Decomposition (SVD)

Another approach we use it in this system which is Singular-Value Decomposition (SVD). The figure (4) explain the form of a SVD. Let X be an \( m \times n \) matrix, and let the rank of X be r. The rank of a matrix is the biggest number of rows “or equally columns” we can select for which number of non-zero linear set of the rows is the vector 0 (all-zero) in this case a set of such columns or rows is independent. Also, the Figure (1.4) displays the matrices U, \( \Sigma \), and V as with the following properties:
Figure 4 The form of a singular-value decomposition

1. $U$ is an $x \times r$ column-orthonormal matrix; that is, each of its columns is a unit vector and the dot product of any two columns is 0.
2. $V$ is an $n \times r$ column-orthonormal matrix, note that we always use $V$ in its transposed form, so it is the rows of $V^T$ that are orthonormal.
3. $\Sigma$ is a diagonal matrix; that is, all elements not on the main diagonal are 0. The elements of $\Sigma$ are called the singular values of $X$.

Singular-Value Decomposition (SVD) algorithm steps are described in the algorithm below:

**Algorithm (020) Singular-Value Decomposition (0SVD0)**

**Input**: Generate Data matrix $X$

**Output**: New Dimensions $C$

1. Repeat
2. Applying SVD to the matrix $X$ as $X = U \Sigma V^T$
   
   $X \rightarrow$ is an $m \times n$ matrix
   
   - $m \rightarrow$ no. of sessions (0vectors)
   
   - $n \rightarrow$ is no. of TH attributes)

   $U \leftarrow XX^T$ matrix 0 of the eigenvectors

   $\Sigma$ is matrix 0 which is diagonal 0

   $V \leftarrow$ is matrix the eigenvectors 0.

3. Construct the covariance matrix $XX^T$ from this decomposition by

   $XX^T \leftarrow (USV^T)(USV^T)^T = (USV^T)(VSU^T)$

4. $V \rightarrow$ an orthogonal matrix ($V^TV = I$), $XX^T = USU^T$

5. Square roots of the eigenvalues of $XX^T$ are the singular values of $X$

6. untilRepresent 0 every transaction $Ii$ over the time interval $t$ as a vector $0x(t)_i$

1. **Return** $U^TX$

4.3.3. **Linear discriminant analysis**

Linear discernment analysis (LDA) is a different technique that employed for reduction of dimensionality and feature extraction. LDA requires reducing dimensionality while maintain as much of the class distinctive information.

LDA algorithm phases are presented in. LDA is a high-dimensional data analysis approach and appropriate for characteristics transformation to ease classifications [22]. There has been capability to utilize PCA method for characteristics subset picking or reduction in many various applications such as face recognition, text recognition and
handwritten, and image compression, in addition to intrusions detection [23] but LDAs has more advantages over PCA's and is favored over PCA's owing to the following aims.

a) LDAs outperforms PCA's in example of great numbers of samples dataset [24].

b) LDAs directly treats with both differentiation within-classes as well as between-classes while PCA does not have any conception of the between-classes structure [25].

c) LDA maintenance class differentiation information as much as possible while accomplishment dimensionality reduction [9].

Algorithm 3 Algorithm for Linear Discriminant Analysis

**LDA Algorithm Steps**

Suppose \( X=(X_1,X_2,X_3, \ldots, X_C) \) are Nx1 features vectors where \( C \) and each features vectors contains \( n \) features. Followings are steps adapted in LDA algorithms.

Steps 0: Computes the between class scatters using complete features samples.

\[
S_b = \sum_{i=1}^{C} (\mathbf{x}_i - \mathbf{a}_i)(\mathbf{x}_i - \mathbf{a}_i)^T
\]

Step 2: Calculates the Total class scatter matrix

\[
S_T = \sum_{i=1}^{C} \sum_{j=1}^{n} (\mathbf{x}_i^j - \mathbf{a})(\mathbf{x}_i^j - \mathbf{a})^T
\]

Step 3: Computes Eigenvalues and Eigenvectors using Eigen equation for LDA.

\( S_T X \rightarrow \lambda X \)

Step 4: Computes the Eigenvectors corresponding to Eigenvalues such that and Eigenvectors: \( X_1, X_2, X_3 \ldots X_N \) where \( N \) represents dimensionality of feature vector and \( N \) in our case

**Eigenvalues:** \( \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots \lambda_N \)

Step 5: Evaluate the contribution of each feature vector

\[
c_j = \sum_{p=1}^{n} |v_{pj}|
\]

Step 6: Sort the features vectors in descending orders corresponding to their impact or contributions.

Step 7: The dimensionality reductions phase based on largest 0 eigenvalues is skipped as the selection of optimum subsets of linear components

4.4. Classification Support Vector Machine Algorithms (SVM)

Support Vector Machine (SVM) is a machine learning method setsemployed for regression and classification. SVM is depending on the idea of decisionplanes that describe decision boundaries. A decision plane is one that splits between a set of matters having various class memberships. Our suggested insertion detection system depend on dimensionality reduction which PCA and SVD, LDA algorithm which has employed one classification outcomes.

4.5. Performance Evaluation

The insertion detection system efficiency is evaluated by its capacity to make precise estimates. According to the real nature of a grant events compared to the forecast from the IDS, four probable results are presented in Table 4,
famous as the confusion matrix [4]. Detection Rates (DR) or True Negative Rates (TNR), True Positive Rates (TPR), False Positive Rates (FPR) or False Alarm Rates (FAR) and False Negative Rates (FNR) are gauges that can be practical to quantify the execution of IDS [4] depend on the above confusion matrix.

Table 4 Confusion Matrix

| Predicted Actuals | 0Negative Class (Normal) | 0Positive Class (Attack) |
|-------------------|-------------------------|-------------------------|
| 0Negative Class (Normal) | True Negative (TN) | 0False Positive (FP) |
| 0Positive Class (Attack) | False Negative (FN) | True Positive (TP) |

We have gotten accuracy by recognition rate is illustrate such as the ratio between the correct recognition numbers decision to the number of total.

\[
\text{Accuracy} = \frac{TP + TN}{\text{Total number of test samples}} \times 100
\]

5. Results and discussion

Tables (5) and (6) display the overall performance results of Support Vector Machines (SVM) on KDDsCup 99 dataset based on testing and training by utilizing three various algorithms (0PCA and LDA, SVD0) that we have offered in our system 0.

Table 5 Accuracy using (SVM) and different feature extraction and dimensionality reduction algorithms on the training dataset

| Dataset0Features | Classification Algorithm | Dimensionality0Reduction Algorithm0s | Accuracy0 |
|------------------|--------------------------|--------------------------------------|------------|
|                  |                          | Algorithm0 | Feature 0No. |            |
| 0420             | SVM                      | PCA        | 07           | 98.43943   |
| 042              | SVM                      | LDA        | 04           | 99.24772   |
| 042              | SVM                      | SVD        | 07           | 98.7197    |

Table 6 Accuracy for using (SVM) with different dimensionality reductions algorithms on the testing dataset

| Datasets0 Features 0 | Classification Algorithm | Dimensionality0Reduction0 Algorithms | Accuracy0 |
|---------------------|--------------------------|--------------------------------------|------------|
|                     |                          | Algorithm0 | Features No. |            |
| 042                 | SVM                      | PCA        | 07           | 91.4771    |
| 042                 | SVM                      | LDA        | 04           | 99.4511    |
| 042                 | SVM                      | SVD        | 07           | 98.0556    |

Figure (5) explains the performance results on the training dataset employing three dimensionality reduction algorithms. For attack detection we utilized support Vector Machine classification algorithm.
5.1. Experiential Results using the Whole Dataset Samples

Big data analysis is a major change these days, so in terms of dealing with a huge number of data samples (records). In our proposed system, we design an approach to handle intrusion detection classification difficult. Also, we used a huge number of data examples (494,201) with entire feature numbers (42 features). To execute and solve the problem of 10% KDD classification, we propose a data folds segmentation. In this part, we tried to display the experimental results employing the entire data samples which are (494,201). These experimental results of intrusion detection proposed system which is the intrusion detection classification system depend on various features reductions algorithms on the KDD Cup 99 dataset.

Each record of dataset labels includes one of the 5 types of attacks. Since the 494,201 is a big data analysis especially in our proposal, we employed three various algorithms to do the dimensionality reduction and feature selection. In each one we used for reducing the 42 features of the 0 KDD data sets and two classification algorithms to detect the four types of IDs attacks.

Our methodology proposed system for dealing with this type of big data analysis is to divide the entire dataset sample (494,201) to k-folds (k-folder). Each fold (folder) has n-sample numbers from the dataset. Each sample has been picked in terms that no fold (folder) has the same data sample as another fold.

In experimental results of proposed system, we divide the dataset to 25-folds. The 24th folds, each one has 20000 data samples, and the last one has 14021 samples so all these shown in Table (7) below.
Table 7 Whole data sample approach

| Total Dataset size | Data approach 25-folds splitting |
|--------------------|----------------------------------|
|                    | 1 to 24th fold | 25th fold |
| 494,201            | 20000 | 14021 |

The experimental results were examined and discussed to exemplify the proposed ID system. In this case, we described three major parts. The first part is the essential features by employing three algorithms that have selected from the entire feature space which is 42 features. The second part, we explained in this step the result of dimensionality reduction imprudent and feature selection algorithm by selecting the feature space (7). The last part is a part of comparing between IDS proposal experimental results and the previous works.

In this implementation system, we trust on scoring the Eigen value score to reorder the feature from the highest score (the most significant one) to the smallest one.

5.2. Classification Experimental Results

classify the kinds of attack on the 10% of KDD Cup 99 dataset we employed. Support Vector Machine (SVM) classification algorithm in this phase. This algorithm has been utilized with the reduction dimension space features.

5.3. Comparing our Classification Results

Compare the performance results of SVM classifiers by employing the whole data samples with all three dimensionality reduction methods that we have suggested.

Table (8) displays the performance results of the SVM employing the whole data sample basing on the utilizing the all dimensionality reduction methods.

Table 8 A Compression results for Insertion Detection using SVM

| Dimensionality Reduction | Dimensionality | Training | Testing |
|--------------------------|----------------|----------|---------|
| PCA                      | 21             | 94.08%   | 93.79%  |
|                          | 11             | 94.10%   | 93.82%  |
|                          | 7              | 94.13%   | 93.83%  |
| LDA                      | 4              | 92.28%   | 98.11%  |
| SVD                      | 21             | 93.36%   | 91.65%  |
|                          | 11             | 93.44%   | 91.75%  |
|                          | 7              | 91.77%   | 90.12%  |

We can see that by employing the SVM classifier to categorize the entire data samples according to the attack kinds, our method for PCA and LDA, SVD gives a higher accuracy in testing and training.
Figure 7 Accuracy Support vector machine (SVM) classification with three Algorithms Dimensionality Reduction Attack Detection

5.4. Comparing our Classification Results with Other studies

There were great number studies that have been done to classify the attack kinds using 10% of KDD Cup 99 dataset. In this part, we will compare our results utilizing reduction algorithm (PCA and LDA, SVD) with the other studies that have been done on the same dataset. Table (9) shows a briefly comparison between our proposed system’s result and the other methods according to the performance results for the overall accuracy for testing and training.

Table 9 A Compression results for Insertion Detection between the previous studies and our approach

| Approach            | Classification Alg. | Feature No. | Accuracy  |
|---------------------|---------------------|-------------|-----------|
| Soni and Sharma (2014) | C5.0               | 32          | 99.49%    |
|                     | ANN                 | 32          | 99.49%    |
| Zargar and Baghaie (2012) | KNN               | 42          | 99.41%    |
|                     | KNN                 | Effective features are used from (42) | 96.70%    |
| Mukkamala and Sung (2003) | ANN               | 41          | 87.07%    |
|                     | ANN                 | 34          | 81.57%    |
| (Zhu et al.,2005)  | BP Network          | 12          | 88.15%.   |
| Ming-Yang Su (2011) | GA KNN              | 19          | 97.42%    |
|                     |                     | 28          | 78.00%    |
| Suggestion          | SVM                 | 7           | 98.00%    |
|                     | SVM                 | 4           | 99.45%    |
6. Conclusion

Today, a large amount of threat attacks network and information security. In this paper, we proposed an intrusion detection system that reduces the set of features and classifies attack types. The reduction of features is performed by us also then the classification which the proposed algorithm is a combination of features selection. Reduced features for intrusive detection system and increased attack detection rate to the SVM applied classification algorithm, which gives the highest resolution. The cup1999 kdd selection attacks are identified with less Error rate and high accuracy. The feature selection and their reduction have both affected the performance of the classification algorithm. In the future swarm optimization function dynamically reduces the number of unused feature attribute of traffic data.

Compliance with ethical standards

Acknowledgments

The authors would like to thank the Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad, Iraq, for supporting this work.

Disclosure of conflict of interest

All authors declare that they have no conflict of interest0.

References

[1] Axelsson, S, "Intrusion Detection Systems: A Taxonomy and Survey," Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.
[2] Lunt, T F, “Detecting Intruders in Computer Systems,” in proceeding of 1993 Conference on Auditing and Computer Technology, 1993.
[3] Sundaram, A, "An Introduction to Intrusion Detection," The ACM Student Magazine, April 1996; Vol.2, No.4. Available at http://www.acm.org/crossroads/xrds2-4/xrds2-4.html
[4] Porras, P A, “STAT: A State Transition Analysis Tool for Intrusion Detection," MSc Thesis, Department of Computer Science, University of California Santa Barara, 1992
[5] Dorothy E. Denning, “An Intrusion Detection Model,” In IEEE Transactions on Software Engineering, February 1987; Vol.SE-13, Number 2: page 222-232.
[6] Lunt, T F, et al., "A Real-time Intrusion Detection Expert System (IDES)," Technical Report SRI-CSL-92-05, Computer Science Laboratory, SRI International, Menlo Park, CA, April 1992.
[7] Mykerjee, B, Heberlein, L T and Levitt, K N, “Network Intrusion Detection,” IEEE Network, 1994; Vol.8, No.3:pp.26-41.
[8] Ilgun, K, Kemmerer, R A, and Porras, P A, "State Transition Analysis: Rule-Based Intrusion Detection Approach," IEEE Transactions on Software Engineering, March 1995; Vol. 21, No. 3:pp.181-199.
[9] Kumar, S, “Classification and Detection of Computer Intrusions,” PhD Thesis, Department of Computer Science, Purdue University, August 1995
[10] A Macgregor, M Hall, P Lorier and J Bruskill, "Flow clustering using machine learning techniques", In PAM 2004, Antibes-Juan-Les-Pins, France, LNCS. 2004; pp. 205-214.
[11] S. Kumar, Classification and Detection of Computer Intrusions, Ph.D. Thesis, Purdue University.
[12] MithcellRowton, Introduction to network security intrusion detection, December 2005.
[13] p Soni , p Sharma, ” An Intrusion Detection System Based on KDD-99 Data using Data Mining Techniques and Feature Selection”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307,July 2014; Volume-4 Issue-3,
[14] G R Zargar and T Baghaie,"Category-Based Intrusion Detection Using PCA",Journal of Information Security, 2012; 3:259-271 http://dx.doi.org/10.4236/jis.2012.34033 Published Online October 2012.
[15] S Mukkamala, A H Sung, "Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines", Transportation Research Record Journal of the Transportation Research Board - January 2003 DOI: 10.3141/1822-05.

[16] Zhu, Y. et al. Modified Genetic Algorithm based Feature Subset Selection in Intrusion Detection System. Proceedings of ISCIT 2005; 9-12.

[17] Ming-Yang Su Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. Expert Systems with Applications, 2011; 38:3492–3498

[18] skDD Cup 1999 Intrusion detection dataset: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[19] Mukkamala, S. et al. (2005). Intrusion detection using an ensemble of intelligent paradigms. Journal of Network and Computer Applications, 28(2), 167–82.

[20] Lindsay I Smith A tutorial on Principal Components Analysis February 26, 2002.

[21] S. Gong et al., Dynamic Vision: From Images to Face Recognition, Imperial College Press, 2001 (pp. 168-173 and Appendix C: Mathematical Details, hard copy

[22] Kresimir Delac, Mislav Grgic and Sonja Grgic, “Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set,” University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia. 2006.

[23] M. Turk and A. Pentland, “Eigen faces for recognition,” J Cogn Neurosci 3 (1991), 71–86.

[24] P. Belhumeur, J. Hespanha, and D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, Proc Fourth EurConf Computer Vision, Vol. 1, 1418 April 1996, Cambridge, UK, pp. 45–58.

[25] A. Martinez and A. Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence," vol. 23, no. 2, pp. 228-233, 2001.

[26] Putra Wanda, Huang Jin Jie "A Survey of Intrusion Detection System" International Journal of Informatics and Computation (IJICOM) Vol.1, No.1, August, 2019 ISSN: 2685-8711

[27] R. Vinayakumar et al. "Deep Learning Approach for Intelligent Intrusion Detection System" Received December 27, 2018, accepted January 3, 2019, date of current version April 11, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2895334

[28] H. Alqahtani et al. "Cyber Intrusion Detection Using Machine Learning Classification Techniques" © Springer Nature Singapore Pte Ltd. 2020 N. Chaubey et al. (Eds.): COMS2 2020, CCIS 1235, pp. 121–131, 2020. https://doi.org/10.1007/978-981-15-6648-6_1