www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 11 Issue 02February 2022, Page **No.25508-25514**
ISSN: 2319-7242 DOI: 10.18535/ijecs/v11i02.4657

Model Based Intrusion Detection using Data Mining Techniques with Feature Reduction

Jyotsna Goyal

Department of Computer Science Engineering and Technology Thapar University Patiala, India

E-mail: jgoyal me16@thapar.edu

Abstract:
As the technology is advancing so are the data storing practices. Nowadays data is stored online which is the main reason as to why the data is constantly under threat. Therefore there is an urgent need of computer security for securing this confidential data, which is mostly customer personal data which if got leaked will not only pose threat to the customer but also to the organization liable for storing and preserving that data. These unwanted activities are termed as intrusions and the detection of these unwanted activities by constantly monitoring and analysing the system is known as intrusion detection. IDS created using data mining techniques is an effective way of detecting intrusions whose implementation is discussed ahead in this paper. The approach involves building of classification model and hybrid model which are created using classification techniques and, combining both classification and clustering techniques respectively. Classification model can detect known attacks effectively whereas hybrid models can detect unknown or new attacks also. NSL-KDD dataset is used as training dataset which is normalalized and then its feature reduction is done using different techniques. The best feature selection technique among all is chosen by using decision table algorithm. The comparison of the results of different models is done over different performance evaluation parameters. The results show that hybrid models perform better than classification models with improved results as the the data is first preprocessed which makes a classifier more effective.

Key words: IDS, Data mining, NSL-KDD, CFS Feature Selection, U2R attacks, R2L attacks.

Introduction

Storing of data and other data-related operations using digital methods provides a lot of benefits like easy data organization and retrieval but also makes data prone to cyber-attacks. In current times the data is transferred over the network in huge amounts which lets the hackers and criminals take advantage of this and misuse the data. The data stored remotely is also critical and is under constant threat from these unauthorized sources. Therefore timely detection and prevention of such activities are very much required everywhere. The unwanted access of data by unauthorized and dangerous criminals is known as intru- sion and the detection of these intrusions by analyzing and constantly monitoring the system which is to be secured from these threats is known as intrusion detection. The intrusion detection systems(IDS) are the systems that can definitely help in the detection of these threats. Intrusion detection includes collecting information that could be con- ferred as an intrusion by monitoring the system and saving logs in the system and then analyzing the output. The data is then characterized as either intrusion or normal behavior. The attacks are of three types that need to be de- tected and can be external attacks (attacks that happen due to access from an element outside the organization) and internal attacks (attacks from one among the users of the organization). These attacks are further divided into misuse-based, anomaly-based, and hybrid attacks. Misuse

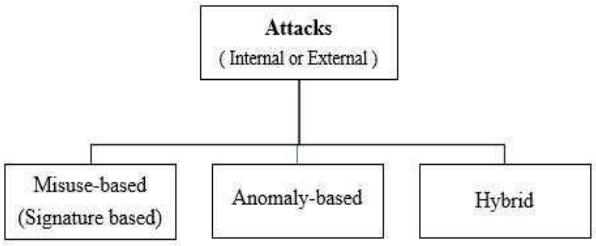

Fig. 1: Types of Attacks

Attacks or signature-based attacks are known attacks that the system has already encountered and therefore already present in the database. These attacks don't generate high false alarm rates but they require frequent updates in the database with rules and signatures. Anomaly-based attacks are unknown attacks that are new to the system. These attacks are also known as zero-day attacks. The main dis- advantage here is that system here gets high false-positive rates (FPR) because unseen but legitimate system behaviors may be categorized as anomalies. Hybrid attacks are in which the system encounters both types of attacks. Intrusion detection models are mainly built using data mining and machine learning methods. Data mining is extracting useful data and patterns from the previously unknown data by applying some specific algorithms. Ma- chine learning on the other hand is making the model learn and then lets to predict the new domains based on known characteristics of the data on which it is built. Data mining is divided into two categories, Supervised learning, and unsupervised learning. Supervised learning is in which the model is constructed from a pre-classified dataset. The dataset contains a class attribute that divides each data point into different sets or classes and new data points are classified into different sets to which they should belong on the basis of some similarity measure whereas unsupervised learning is a method that applies to the data which is not pre-classified, so the classification is done on the basis of clustering. The data points are classified into different clusters on the basis of similarity between the data points and the mean points for each cluster. The new data points are then classified into the cluster to which they most appropriately belong. Sometimes both classification and clustering methods need to be combined for detecting both known and un- known attacks which contribute to hybrid models.

Literature Review

In the literature, several data mining and machine learning techniques have been applied for intrusion detection system (IDS). Ng et al. [1] presents a tool which is used for detecting both anomaly detection and a signature based detection using a log file to detect patterns that may be considered an unauthorized activity. If the pattern is con- sidered as an attack it is stored in the database for future use. The tools keeps on learning and gets more powerful, the new attacks are detected and classified using clustering which includes grouping similar activities together based on popular trends. The concept of reoccurring matches with the DoS attack which enables detection of possible password guessing. Singh et al. [2] used different data mining algorithm for finding correct patterns of an attack. Zhu et al. [3] intrusions are detected by first recording host logs and then intrusions are detected using methods like ARIMA time series, Apriori association algorithm. Then, intrusion detection methods which are misuse-based and anomaly -based methods are integrated and applied to host system and through this the system security is enhanced. The system is unable or has some issues for detecting unknown or zero day attacks but the detecting of data from logs is done with accuracy. Their methods also coped with the rapid increase in the size of log files thus improving the system stability. Puthran et al. [4] has used different data mining methods to detect and analyze different attacks using methods like classification and clustering. They have mainly focused on the attacks prevailing in the network. They concluded that attacks are detected accurately when using both types of mining techniques. Shahadat et al. [5] used decision Table to detect intrusions. The decision table is a rule based mining technique. This algorithm gave better performance and accuracy over the existing models. To select important features they used a new technique called Dropout (DP). which does sequential search and drops out all non-relevant features keeping only the important features. Sultana et al. [6] contributed in detection of net- work intrusion using Average one dependence estimators (AODE) which according to the research is an improvement over the na¨ıve bayes algorithm. This model first create some estimators and then averages the results produced. This algorithm helps in detecting different types of attacks and gave efficient results. Gupta et al. [7] used techniques like K-Means Clustering, Linear

Regression generate rules automatically and detect intrusions based on these rules. Zhao et al. [8] presented work which analyzes the audit data by extracting the properties deeply and then these intrusion characteristics of the network are analyzed and then combined intrusion detection techniques along with human intervention to establish a rule base using C.45 algorithm. This paper used optimal pattern matching algorithm so that the detection rate can be improved. Kathleen et al. [9] used a model formed by using Na¨ıve Bayes, decision trees and SVM as base classifiers This model showed high accuracy while reducing false positive rates. Firstly, they used SVM to divide the dataset into normal and attack, then attack points is fed to a decision tree and Naive Bayes which is used to classify these attacks.

Research Approach

NSL-KDD Dataset

The NSL-KDD dataset is used for preparing intrusion de- tection model which is a refined version of KDDcupp99 dataset. The work includes different tools and techniques used to develop an effective intrusion detection model on the NSL-KDD dataset. This dataset consists of training and testing dataset. The dataset includes 41 attributes and a label class which suggests the attack type as either normal or one of the attacks. The first nine features are basic features of network connection vector, next thirteen features are content features of network connection vector, and next nine are time based network traffic features and rest i.e. the last ten are host based network traffic features.

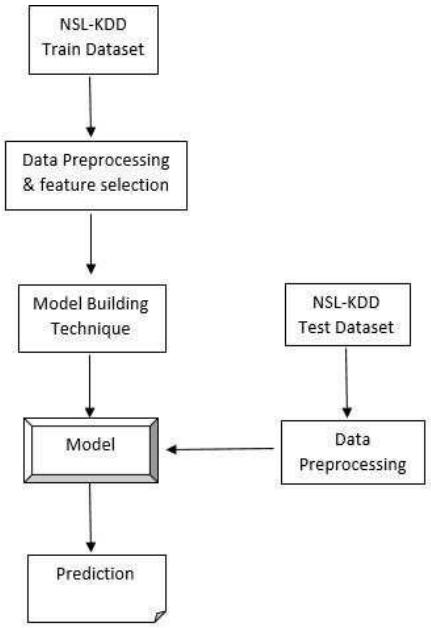

Fig. 2: Basic Work Approach

The attack class present in the NSL-KDD training dataset contribute to four type of attacks:

1) DOS: Denial of service is a type of attack in which any attacker tries to make a network or any resource unavailable to the legitimate users by disrupting the services of a host. e.g. flooding the targeted machine or resource.

2) Probe: Probe is a type of attack in which the attacker aims at accessing the information about the remote system or tries to know the network state. e.g. A user trying to send an empty e-mail just to check if the person is online or to know the target person's usage patterns.

3) U2R: U2R is any unwanted access to exploit the vulnerabilities in the victim's system by accessing its local account to get the root/admin privileges of the system, mainly in an organisation e.g. kernel level attacks

4) R2L: R2L is any unwanted access from a remote machine to a local machine(remote to local). In this type of attack the attacker tries to get into a remote machine to gain local access of the victim machine. E.g. remotely logging into a system by knowing its username and password. The normal class label tells that the activity does not account to any of the attacks. The test set contain many other attacks which are not present in training dataset.

Preprocessing

The raw data is comprised of data with varying scales. Nor- malizing the dataset will give a boost to the performance of the classifiers. The normalization of dataset can be done using feature scaling, z-score and many other techniques. We have normalized dataset using Feature scaling which gave better results than z-score. Rescaling or min-max scaling is the most simple method of all and here the values of the features are converted to a scale of [0, 1] or [−1, 1] and is as below:

:

$$f' = \frac{f - min(f)}{max(f) - min(f)} \quad (1)$$

f is an original value,
f' is the normalized value

Feature Selection

Feature selection is done as it is required for dimensionality reduction and removal of unimportant features which in turn also improves the accuracy and other parameters related to the outcome. The feature selection techniques like CFS (Correlation Feature Selection) subset evaluation with greedy search, Pearson correlation, Info gain attribute evaluation and Info gain with filter method techniques were used. We have created our approach in which, the feature set is first applied with ranking algorithm which ranks the features according to the information gained in descending order and then applied with decision table algorithm which gives accuracy of the model in prediction. Then, the features from the end are removed one by one till the accuracy us- ing decision table remains constant. The Feature Reduction technique results are as below- The results from all the techniques are evaluated using decision table algorithm.

TABLE 1: Feature Selection Techniques

Selection Technique	Count	Features Selected	Accuracy
CFS+greedy search	11	3,4,5,6,12,14,26, 29,30,37,38,42	98.82%
Pearson Correlation	12	3,4,5,6,12,14,26, 29,30,37,38, 14,42	98.81%
Info gain evaluation	30	1,2,3,4,5,6,8, 10,12,22,23,24, 25,26,27,28,29, 30,31,32,33,34, 35,36,37,38,39, 40,41,42	98.26%
Our approach	23	2,3,4,5,6,12,23, 24,25,26,29,30,31, 32,33,34,35,36, 37,38,39,40,42	98.27%

The results show that Pearson correlation technique has come out to be most accurate so the respective feature set will be used further. The new feature set is given ahead as shown in Table 2.

Table 2: New Feature Set

Feature No.	Feature Name
1.	Service
2.	Flag
3.	Src bytes
4.	Dst bytes
5.	Logged in
6.	Root shell
7.	Srv serror rate
8.	Same srv rate
9.	Diff srv rate
10.	Dst host srv diff host rate
11.	Dst host serror rate
12.	Class(normal or attack)

Model Building Techniques

Classification Algorithm

The model built using classification techniques are more appropriate in detecting signature based attacks i.e. attacks which are already present in the database. They are not as efficient in detecting novel attacks. The approach is shown in Figure 3.

Hybrid Algorithm

The model is built using both classification and clustering methods. Many new type of attacks remain undetected using classification techniques so the new type of attacks are handled using clustered data and the signature attacks are detected and classified using the classification data. The approach is shown in Figure 4.

CLASSIFICATION ALGORITHMS

A. C4.5

C4.5 algorithm is used to generate a decision tree, which is an improvement over the earlier decision tree algorithm

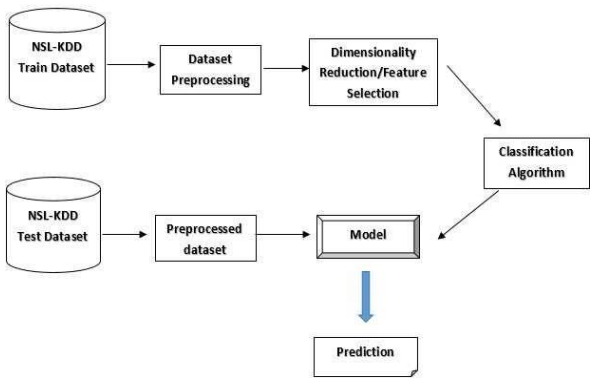

Fig. 3: Work Mechanism for classification model

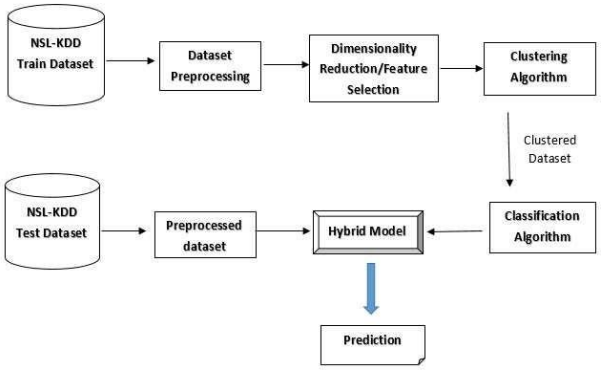

Fig. 4: Work Mechanism for hybrid model

ID3. C4.5 algorithm is a statistical classifier which classifies the nodes on the concept of information gain. The training data is set of classified samples having p-dimensional vec- tors defining the attributes of the samples. C4.5 generates a decision tree where each class at each step is selected on the basis of maximum information gain i.e. the node with the maximum information gain is selected as the splitting criteria.

K-nearest neighbors (KNN)

KNN is a an algorithm which plots whole test sample into n-dimensional space and then finds the k-nearest neighbors using any of the distance measures (like Manhattan dis- tance, Euclidean distance) from this space. The learner do this whenever it needs to predict for a test sample and it goes through the full training dataset every time for this step, so it is called a lazy learner also.

Ripper

Ripper is a rule-based learner which construct set of rules for classifying the problem. The learner does not make any a priori assumptions to reach the final concept rather it works on the assumption that the data on which it is trained is similar in way as the unseen data. This algorithm generate a set of hypothesis that more appropriately generate the target concept.

Naive Bayes

Naive Bayes algorithm use probability theory and the Bayes theorem to predict the category of a sample. Here, the probability of each category is calculated for a given sample based on the prior knowledge of the conditions that might be related to that feature.

Random Forest

Random forest is a supervised learning algorithm which creates a forest of number of trees for classifying the prob- lem. Higher the number of trees, better is the accuracy of the model. In random forest instead of choosing the root node by using the Gini index or information gain, the root node and the further nodes are chosen randomly. This lets in creation of different decision trees and then these trees contribute to the random forest.

Random Subspace

The Random subspace method is also called attribute bag- ging or feature bagging method as it is an ensemble learns- ing method and uses bagging as an ensemble technique. Here good classifiers are combined to form an efficient model. The individual learners are combined by using majority voting or posterior probabilities. We have used adaboost and decision tree algorithms as base learners.

Results and Analysis

From the table (Table 3) it can be seen that Ripper and C4.5 out performs among all the techniques giving a good accuracy score with the former giving highest true positive rate and latter giving lowest false positive rate.

From the graph it can be visualized that KNN is giving better results for the attacks U2R and R2L as compared but these attacks constitute very lesser part of the complete dataset which nullifies the effect of giving the best perfor- mance for these attacks thus Ripper and C4.5 out performs among all and can be seen from the table as well as figure above.

TABLE 3: Performance Evaluation of Classification Algorithms

Classify Algorithm	TPR	FPR	Precision	MCC	Accuracy
C4.5	0.855	0.113	0.859	0.750	85.54%
KNN	0.843	0.117	0.858	0.735	84.27%
Ripper	0.857	0.136	0.881	0.732	85.67%
Naive Bayes	0.787	0.183	0.803	0.646	78.72%
Random Forest	0.839	0.137	0.843	0.711	83.89%
Random Subspace	0.837	0.141	0.905	0.5864	83.70%

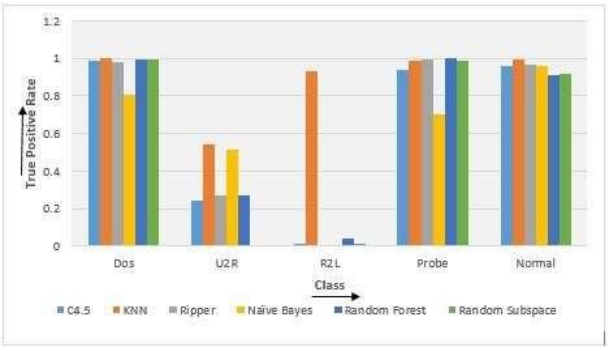

Fig. 5: Analysis of Classification Algorithms for different class labels over True Positive Rate

Hybrid Algorithms:

We have used K-means as clustering algorithm along with four of the above discussed classification algorithm for building four different hybrid models. K-means algorithm is a method in which the algorithm creates k clusters and distribute the data points into these clusters on the basis some distance measure. Then the new data point is clustered in the appropriate cluster by first finding its distance from the mean value of each cluster and then classifying it in the cluster with least distance.

Kmeans + C4.5

First K-means is applied on the dataset and then, C4.5 algorithm is applied on the clustered dataset.

K-means + Naive Bayes

Firstly, the K-means algorithm is applied on the dataset which is applied in the same way as above and then the naive Bayes algorithm is applied on the clustered dataset.

K-means + Ripper

The dataset is firstly passed through the K-means clustering algorithm and then the Ripper classification algorithm is applied.

K-means + Random Forest

After applying K-means algorithm to the dataset, random forest is used for further classification.

Result and Analysis

From the above table (Table 4) it can be seen that all the techniques give fairly comparative result and give good accuracy on unknown data also. K-means with C4.5 gives the highest true positive rate and K-means with Nave Bayes gives the lowest false positive rate. The graph depicts that K-means with Naive Bayes im- proves the true positive rate of every attack and also gives a good accuracy. However other three techniques still give better results as U2R and R2L contribute to very lesser part of the dataset and these techniques have higher true positive rate as compared to K-means with Naive Bayes for other more likely attacks.

Table 4: Performance Evaluation Of Hybrid Algorithms

Hybrid Algorithm	TPR	FPR	Precision	MCC	Accuracy
K-means + C4.5	0.859	0.130	0.854	0.743	85.11%
K-means + Naive Bayes	0.845	0.087	0.890	0.779	84.51%
K-means + Ripper	0.848	0.146	0.887	0.778	84.81%
K-means + Random Forest	0.827	0.143	0.831	0.682	82.73%

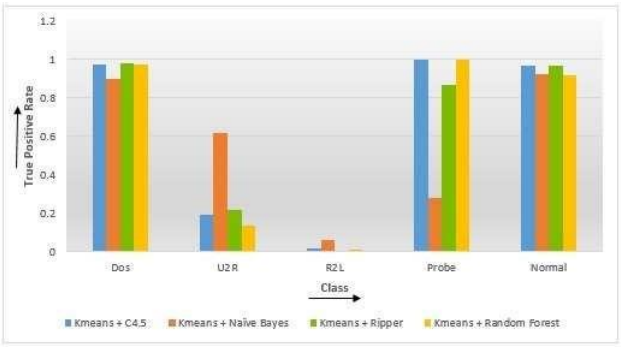

Fig. 6: Analysis of Hybrid Algorithms for different class labels over True Positive Rate

Conclusion:

From the results of classification models and hybrid models it can be concluded that hybrid models perform much bet- ter as they detect new types of attacks(unknown) whereas, the classification algorithm ignore the new attacks. Hy- brid models give better true positive rate and lower false positive rate compared to classification models and use the advantage of classification algorithm by maintaining lower false positive rate for known attacks as they are implemented using classification algorithms as well. On comparison between the results of classification mod- els it can be concluded that among all the classification models C4.5 model has given the best performance. On comparison between the results of hybrid models it can be concluded that among all the hybrid models, K-means with C4.5 model has given the best performance. C4.5 classifica- tion model has

given an accuracy of 85.54% and K-means with C4.5 hybrid model has given an accuracy of 85.11%. Rest of the models have also shown a good performance as can be seen from the results. The results have shown improved accuracy as compared to earlier works using same dataset as we have normalized the dataset and have done feature selection, thus making the classifiers more effective.

Future Scope

The dimensionality reduction of dataset using principal component analysis can be achieved for better results and better data visualization. More advanced algorithms like neural networks can be used to model the system for better detection rate. The system can be made to respond to intrusions along with the detection with lesser false alarms.

References:

[1] Jonathon Ng, Deepti Joshi, and Shankar M Banik. Applying data mining techniques to intrusion detection. In Information Technology-New Generations (ITNG), 2015 12th International Conference on, pages 800–801. IEEE, 2015.

[2] Varsha Singh, Shubha Puthran, and Avanish Tiwari. Intrusion detec- tion using data mining with correlation. In Convergence in Technology (I2CT), 2017 2nd International Conference for, pages 620–625. IEEE, 2017.

[3] Ming Zhu and ZiLi Huang. Intrusion detection system based on data mining for host log.

[4] Varsha Singh and Shubha Puthran. Intrusion detection system using data mining a review. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Confer- ence on, pages 587–592. IEEE, 2016.

[5] Nazmul Shahadat, Imam Hossain, Anisur Rohman, and Nawshi Matin. Experimental analysis of data mining application for intrusion detection with feature reduction. In Electrical, Computer and Commu- nication Engineering (ECCE), International Conference on, pages 209–216. IEEE, 2017.

[6] Amreen Sultana and MA Jabbar. Intelligent network intrusion detec- tion system using data mining techniques. In Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd Interna- tional Conference on, pages 329–333. IEEE, 2016.

[7] Dikshant Gupta, Suhani Singhal, Shamita Malik, and Archana Singh. Network intrusion detection system using various data mining tech- niques. In Research Advances in Integrated Navigation Systems (RAINS), International Conference on, pages 1–6. IEEE, 2016.

[8] Yanjie Zhao. Network intrusion detection system model based on data mining. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2016 17th IEEE/ACIS Inter- national Conference on, pages 155–160. IEEE, 2016.

[9] Kathleen Goeschel. Reducing false positives in intrusion detection sys- tems using data-mining techniques utilizing support vector machines, decision trees, and naive bayes for off-line analysis. In SoutheastCon, 2016, pages 1–6. IEEE, 2016.