JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS IN INTRUSIONDETECTION

KANAKALA SWATHI[1], GUDIWAKA VIJAYA LAKSHMI[2],
M. Tech Student, Department of Computer science and Engineering[1], Assistant Professor, Department of Computer science and Engineering[2]
Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, INDIA[1]

***ABSTRACT*: —** In the earlier many years, the quick improvement of Intrusion Detection and Prevention frameworks assumed an urgent part in PCs organization and security. Interruption recognition framework of Intrusion detection system (IDS) is one of the executed arrangements against hurtful assaults. Moreover, assailants consistently continue to change their apparatuses and methods. Notwithstanding, carrying out an acknowledged IDS framework is additionally a difficult errand. In this paper, a few investigations have been performed and assessed to survey different managed learning classifiers dependent on KDD interruption dataset. It prevailed to figure a few presentation measurements to assess the chose classifiers. The attention was on bogus negative and bogus positive execution measurements to upgrade the recognition pace of the interruption discovery framework. The executed analyses showed that the KNN Classifier accomplished the most accurate results in determining the intrusion rate when compared to the other algorithms like SVM (support Vector Machine), DT (Decision Tree), LR (Logistic Regression) and GNB (Gaussion naïve Bayes). So, application of KNN classifier while designing the system will give us results in a good way.

Key Words: Intrusion detection, KDD dataset, Supervised Machine learning Algorithms, Accuracy, Performance metrics.

INTRODUCTION: There are many sorts of assaults compromising the accessibility, trustworthiness and secrecy of PC organizations. The Denial of administration assault (DOS) considered as perhaps the most well-known destructive attack. An organization can be secured against such assaults utilizing an interruption location framework. An IDS framework can distinguish interruptions and interruption de-produces a ready when it identifies an interruption. This interruption location framework in an organization examinations all traffic. For huge datacenter's this is a troublesome undertaking. There's a huge measure of information through the organization of a datacenter. Standard interruption frameworks can't then all traffic totally.

Interruption identification framework in Intrusion Detection System (IDS) have been acquainted as a device planned with upgrade the security of frameworks [1]. Different IDS approaches have been proposed in the writing since beginnings, however two of them are proposed by Steniford at al., and Denning are generally pertinent in this specific situation. Denning's proposition for an interruption discovery framework zeroed in on the best way to foster successful and exact strategies for interruption location.

As a rule, there are two kinds of IDS (inconsistency base or abuse base). Irregularity interruption location framework carried out to identify assaults dependent on recorded ordinary conduct. Accordingly, it contrasts the current constant deals and past recorded typical continuous deals, this kind of interruption identification framework is generally utilized on the grounds that it can identify the new sort of interruptions. Be that as it may, according to another viewpoint, it enrolls the biggest upsides of bogus positive caution, which implies there is countless typical parcels considered as assaults bundles. In any case, abuse interruption discovery framework is executed to recognize assaults dependent on storehouse of assaults marks. It has no bogus caution and yet, the new kind of assault (new signature) can prevail to go through it.

Attack's detection considered as classification problem because the target is to clarify whether the packet either normal or attack packet. Therefore, the model of accepted intrusion detection system can be implemented based on significant machine learning algorithms. In this paper, the following implemented the machine learning algorithms have been Implemented (Support vector Machine, Decision Tree, GNB classifier, Logistic Regression, and K- Nearest Neighbor) to evaluate and accurate the model of intrusion detection system based on a bench market dataset

Knowledge Discovery in Databases (KDD) which includes the following types of attacks (DOS, R2L, U2R, and PROBE).

RELATED WORK

As a rule, this paper orders IDSs based on recognition strategies they utilize into two classes, as

(i) misuse recognition and

(ii) irregularity identification.

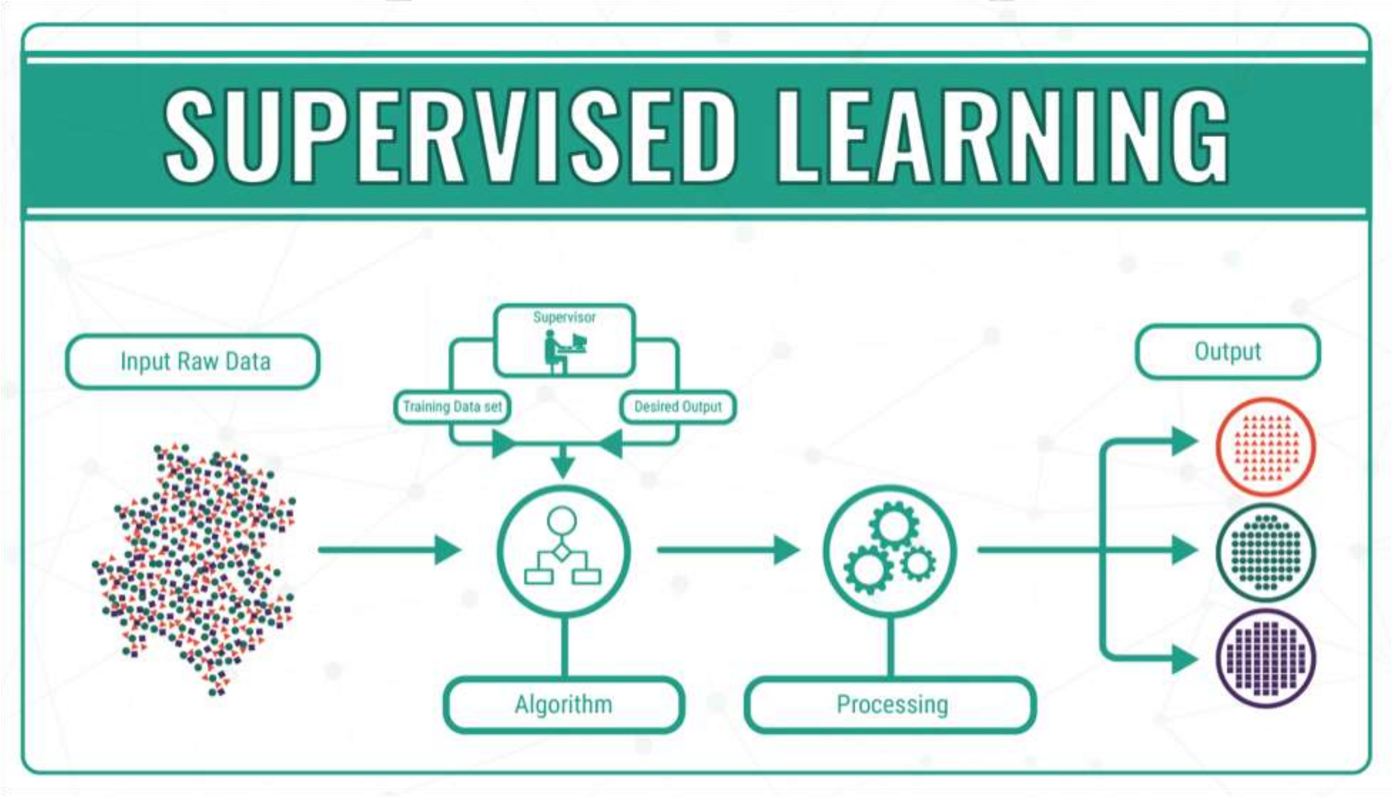
By coordinating with noticed information, abuse recognition recognizes interruptions with pre-characterized portrayals of meddling conduct. So conspicuous interruptions can be recognized in a proficient way using a low bogus positive rate. Accordingly, this strategy is generally taken on in most of business frameworks. Be that as it may, the sorts of new interruptions have developed each second and persistently, in this manner. Abuse the past methods for interruption recognition will neglect to recognize new obscure interruptions. The best way to get freed over this issue is to gain from all interruptions and get update the information at each second. This refreshing system can either be manual or programmed, the manual interaction may be extremely tedious and furthermore the human intercession is needed at each snapshot of time. This interaction can work consequently utilizing administered AI procedures. Sadly, the readiness of datasets for preparing the directed learning calculations is extremely challenging and costly, as this requires assortment and naming of every occasion as ordinary or an interruption type.

This section presents the related works relevant to using KDD dataset for implementing machine learning algorithms. It also provides a brief overview of the different machine learning algorithms and shows how the KDD dataset is very useful for evaluating and testing various types of machine learning algorithms. As per model proposed authors imported the KDD dataset and implemented the pre-process phases e.g. normalization of the attributes range to [-1, 1] and converting symbolic attributes. Neural network feed forward was implemented in two experiments. The authors have concluded that neural network is not suitable enough for R2L and U2R attacks but on the other hand, it was recorded acceptable accuracy rate for DOS and PROBE attacks. As it relates to implement neural network against KDD intrusions, the effort of [8] the authors succeeded to implement the following four algorithms: Fuzzy ARTMAP, Radial-based Function, Back propagation (BP) and Perceptron-back propagation-hybrid (PBH). The four algorithms evaluated and tested for intrusions detection the BP and PBH algorithms recorded highest accuracy rate.

From another study, some of the authors focus on attributes selection algorithms in order to reduce the cost of computation time. In [9] the authors are focused on selecting the most significant attributes to design IDS that have a high accuracy rate with low computation time. 10% of KDD was used for training and testing. They implemented detection system based on extended classifier system and neural network to reduce false positive alarm as much as possible. On the other hand in the information gain algorithm was implemented as one of effective attributes selection. They implemented multivariate method as linear machine method to detect the denial-of-service intrusions.

SUPERVISED LEARNING TECHNIQUES

Machine learning algorithms are sorted as Supervised, Unsupervised, Semi-administered learning and Reinforcement learning. The principal parts of administered learning are forecast, order and relapse. Directed can apply what has been realized in the past to new information utilizing named guides to foresee future occasions beginning from the examination of known preparing dataset, the learning calculation delivers a construed capacity to make expectations about the yield esteems. The arrangement alludes to foresee the discrete esteemed yield i.e., 0 or 1. Though relapse predicts ceaseless esteemed yields. In arrangement we are attempting to plan input factors into discrete classifications and in relapse we attempt to plan input factors to some consistent capacity.



Supervised learning categories and techniques

- Linear classifier (numerical functions)
  Logistic regression, SVM, MLP etc.,
- Parametric (probabilistic functions)
  Naïve Bayes, Gaussian Discriminant analysis etc.,
- Non-parametric (Instance based functions)
  K-nearest neighbour, kernel regression etc.,
- Non-metric (symbolic functions)
  Classification and regression tree, decision tree etc.,
- Aggregation
  Ada boost, Random forest etc.,

In this paper we select an algorithm from each above category and build a model for the prediction and accuracy, precision and other metrics. The selected algorithms are as follows.

Random forest Classifier: is one of tree classifiers utilizing this arranging the quantity of trees ought to be fixed prior to executing. Every individual tree addresses a solitary choice tree. Every individual tree has arbitrarily chosen ascribes from dataset. Hence, the irregular tree classifier could be considered as a limited gathering of choice trees. The methodology of foreseeing the whole dataset is to move a few choice trees yields and pick the champ expected class dependent on absolute quantities of votes.

K - Nearest Neighbor (kNN) is an extremely straightforward, straightforward, flexible and one of the highest AI calculations. KNN utilized in the assortment of uses like money, medical care, political theory, penmanship discovery, picture acknowledgment and video acknowledgment. In Credit evaluations, monetary foundations will anticipate the FICO score of clients. In credit dispensing, banking organizations will foresee whether the advance is protected or unsafe. In political theory, grouping possible electors in two classes will cast a ballot or will not cast a ballot. kNN calculation utilized for both arrangement and relapse issues. kNN calculation dependent on highlight likeness approach.

Decision Trees algorithm is so easy compared with other classification algorithms. Tree representation is used to solve a problem using Decision Tree. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

The Gaussian Processes Classifier or GNB is a classification machine learning algorithm. Gaussian Processes are a generalization of the Gaussian probability distribution and can be used as the basis for sophisticated non-parametric machine learning algorithms for classification and regression.

Logistic regression assumes a Gaussian distribution for the numeric input variables and can model binary classification problems. You can construct a logistic regression model using the Logistic Regression class.

Support Vector Machine is a Supervised machine Learning algorithm which can be used for classification and regression. It implements using the Hypertext plane for classification.

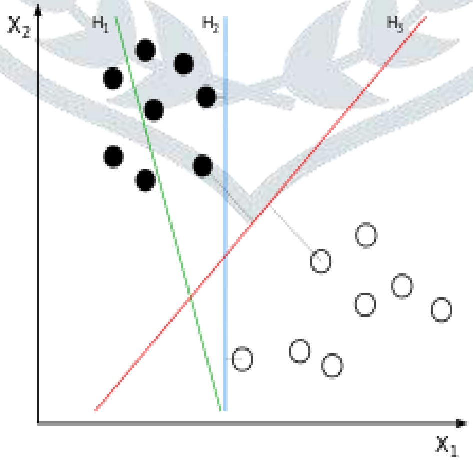


Figure 1 SVN Classification

ABOUT THE DATASET

NSL – KDD intrusion detection dataset is used by many researchers to build an effective network intrusion detection system during past years. But the recent study illustrates there are some limitations are present in the KDD CUP 1999 dataset. It shows that the dataset has 78% training and 75% testing records having redundant records, it inhibits the IDS from detecting rare attacks such as U2R and R2L. The new dataset NSL-KDD dataset is the advanced version of KDD CUP 1999 dataset. Now it is commonly used by the researchers to apply their experiments for analyse the intrusions. The Duplicate records are eliminated to produce an unbiased result. Anming (GEP). The DARPA 1998 dataset was used and 24 attacks can be classified into four types. Adequate numbers of records are available in the train and test datasets to execute the experiments completely. In each record there are 41 attributes are available and a label is present to categorize the data either as normal or an attack type. The attack classes grouped into four types as its predecessor.

Statistics of redundant records in the KDD train set (Original records | Distinct records | Reduction rate)

- Attacks: 3,925,650 | 262,178 | 93.32%
- Normal: 972,781 | 812,814 | 16.44%
- Total: 4,898,431 | 1,074,992 | 78.05%

Statistics of redundant records in the KDD test set (Original records | Distinct records | Reduction rate)

- Attacks: 250,436 | 29,378 | 88.26%
- Normal: 60,591 | 47,911 | 20.92%
- Total: 311,027 | 77,289 | 75.15%

EXPERIMENT RESULT

Performance metrics:

These metrics are used calculating and observing the performance of the IDS and for comparing the results obtained from the dataset. The performance of the intrusion detection system (IDS) is evaluated by calculating four metric values, Accuracy, Precision, Recall and F-score, out of which accuracy plays a major role and the performance evaluation of the IDS is mainly dependent on accuracy metric.

1.Accuracy: This metric is calculated by finding the total number of instances that are correctly predicted as positive cases to the total number of data that is present, the instances are classified into positive or negative cases by calculating the data that are divided into True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN).True Positives(TP) are the data which are correctly classified as true instances, True Negatives(TN) are the data which are correctly classified as false instances, False Positives (FP) refers to the data that are negative instances but are predicted as positive and False Negative(FN) refers to the data that are positive instances which are predicted as negative. The accuracy rate at the maximum times can be taken as high though there are less number of negative instances which does not play a major role in decreasing the accuracy rate, it is calculated as: Accuracy=TP+TN/TP+TN+FP+FN

2.Precision: Precision refers to the total data which are correctly predicted to be positive over the total number of data that are predicted to be positive, by observing the false positive and true positive instances, precision can be calculated as:

Precision = TP / TP+FP

3.Recall: Recall also known as a True Positive Rate (TPR), sensitivity (SN) or detection rate indicates the total number of instances that are correctly predicted as positive over the total number of actually positive instances present. While detecting the overall positive data in the dataset the recall serves as the main evaluation metric or the best performance indicator of positive data, it is calculated as follows:

Recall = TP / FN+TP

Precision and Recall are equally important for calculating the performance of the IDS, each individual is not sufficient for the evaluation of the performance of IDS.

4.F-score: F-score is calculated by considering both the metrics of precision and recall equally, the f1 and f2 scores are calculated, in case of f1 both the metrics are treated equally and the value is obtained by substituting 1 in the place of f-beta, in the case of f2 score the recall is considered two times more important than precision.

Table. Supervised learning technique and its accuracy

Supervised learning technique	Accuracy
Logistic regression	98.02%
SVM	98.07%
K-Nearest neighbor	98.64%
Decision tree	98.55%
GNB	97.88%

From the table it was clear that the KNN classifier algorithm had more accuracy when compared to other algorithms where the GNB algorithm had the least accuracy.

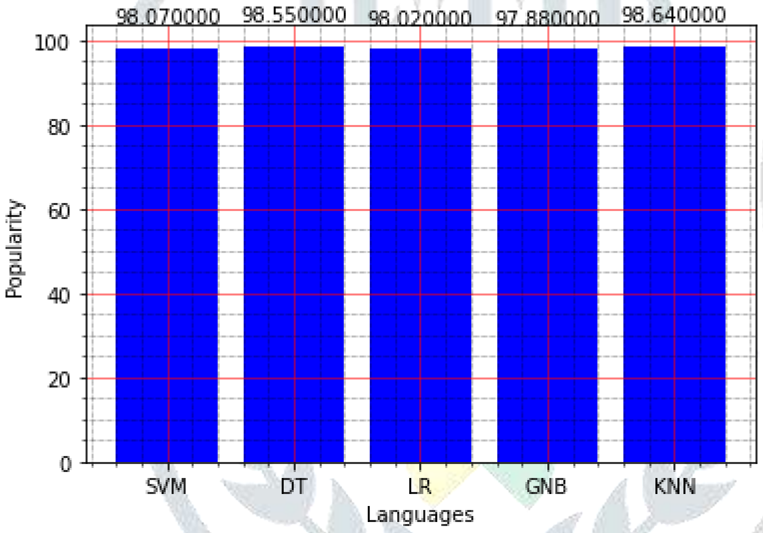


Figure 2 Performance Metrics

As in the graph accuracy is taken on one axis and algorithms are on other axis and compared where KNN classifier is shown as the best.

CONCLUSION

An intrusion detection system is a tool used for automatic detection and removal of external attack or access to the system and takes a decision to determine whether these attacks constitute a legitimate use of the system or are intrusions. So, through the observations which are made it is observed that KNN classifier is the best in accuracy in knowing the intrusions. So, if that algorithm can be embedded then results will be great in performance.

REFERENCES

[1] G. C. Kessler, "Defenses against distributed denial of service attacks," SANS Institute, vol. 2002, 2000.View publication stats

[2] H. A. Nguyen and D. Choi, "Application of data mining to network intrusion detection: classifier selection model," in Asia-Pacific Network Operations and Management Symposium. Springer, 2008, pp.

399–408.

[3] S. Paliwal and R. Gupta, “Denial-of-service, probing & remote to user (r2l) attack detection using genetic algorithm,” International Journal of Computer Applications, vol. 60, no. 19, pp. 57–62, 2012.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on. IEEE, 2009, pp. 1–6.

[5] P. Amudha, S. Karthik, and S. Sivakumari, “Classification techniques for intrusion detection-an overview,” International Journal of Computer Applications, vol. 76, no. 16, 2013.

[6] F. Haddadi, S. Khanchi, M. Shetabi, and V. Derhami, “Intrusion detection and attack classification using feed-forward neural network,” in Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE, 2010, pp. 262–266.

[7] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, “Hide: a hierarchical network intrusion detection system us ing statistical preprocessing and neural network classification,” in Proc. IEEE Workshop on Information Assurance and Security, 2001, pp. 85–90.

[8] W. Alsharafat, “Applying artificial neural network and extended classifier system for network intrusion detection.” International Arab Journal of Information Technology (IAJIT), vol. 10, no. 3, 2

[9] M. Alkasassbeh, G. Al-Naymat, A. B. Hassanat, and M. Almseidin, “Detecting distributed denial of service attacks using data mining techniques,” International Journal of Advanced Computer Science & Applications, vol. 1, no. 7, pp. 436–445.

[10] S. D. Bay, “The uci kdd archive [http://kdd. ics. uci. edu]. irvine, ca: University of california,” Department of Information and Computer Science, vol. 404, p. 405, 1999.

[11] M. Al-Kasassbeh, “Network intrusion detection with wiener filter-based agent,” World Appl. Sci. J, vol. 13, no. 11, pp. 2372–2384, 2011.

[12] S. K. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, and classification,” IEEE Transactions on neural networks, vol. 3, no. 5, pp. 683–697, 1992