ISSN 2349-4506
Impact Factor: 2.265

Global Journal of Engineering Science and Research Management

# IMPROVE THE ACCURACY OF CLASSIFIERS PERFORMANCE USING MACHINE LEARNING & DATA PREPROCESSED METHODS ON NSL-KDD DATA SETS

**Mr. Shobhan Kumar*, Mr. Naveen D.C**
* Computer Science & Engineering Dept. NMAMIT Nitte Udupi(D) Karnataka

**DOI**: 10.5281/zenodo.53749

**KEYWORDS:** IDS, KNN, SVM, VT, NSL_KDD.

## ABSTRACT

Classification is the method of discovering a set of models that describes data classes for the purpose of being able to utilize the model to forecast the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data set. Since the class label of each training sample provides this step is referred as supervised learning. The manuscript describes a system that uses Feature Selection [17, 18] as a data pre-processing activities. Feature selection may present us with the means to reduce the number of network parameters made while still maintaining or even elevating the accuracy and reducing false negative rates. In this document, we used variance-Threshold method which finds an optimum feature subset that enhances the classification accuracy. After this step various classifiers are used such as support vector machine, KNN [12]. Experiments were conducted on the NSL_KDD dataset to assess the effectiveness of our approach. The results prove that SVM Ranking with variance threshold feature selection approach leads to promising step up to feature selection and enhances classification accuracy. Based on the system output the Accuracy and error rate of each classifier is computed.

## INTRODUCTION

Since the Internet is regulated, unmanaged and uncontrolled, it introduces a wide range of risks and threats to the systems operating on it. This is the motive that the network intrusion detection systems (NIDSs) [15] have been emerging recently. Due to the large set of users on Internet, Intranet and extranet network computer access, interference into computer systems by illegal users is an increasing problem. An intrusion is illegal access or attempted access into or unauthorized movement in a computer or information system. Intrusion discovery techniques are therefore becoming extremely important to advance the overall security of the workstation. Intrusion detection is the practice of identifying that an intrusion has been attempted, is occurring or has occurred.

In general intrusion recognition systems, data may be automatically collected and condensed but the analysis of that data usually remains manual. The off-line investigation involves determining normal behavior for a user, application or system. The standard behavior is then used to build a set of rules. The noteworthy deviations from the rules, referred to as anomalous behavior, may then be flagged as possible intrusions. Some intrusion discovery models, based on anomaly detection techniques, look for statistically anomalous behavior, that is, behavior that appears unusual when compared to other user behavior.

One negative aspect of anomaly revealing models is that they are prone to both false positive and false negative alerts since the rules are general in nature and not specific to the activities of each user. False positives take place when the intrusion detection system recognizes an incident as an intrusion when none has occurred. False positives may reroute the thought and time of the system administrator and security staff and if recurrent enough, may cause a lack of assurance in the intrusion detection models. False negatives are instances where the intrusion detection system fails to distinguish an intrusion while it is occurring or after it has occurred. The effect may be slow or no response to the intrusion that can outcome in monetary loss and system damage. False negatives often occur, since the models used to profile the anomalous behavior do not adequately forecast the intrusive behavior and its result within the computer system. The major drawbacks of IDS [15] are that the systems can only detect and monitor what has been previously defined to them, either using proficient system rules or rules developed through data collection, dimensionality reduction and analysis. This can outcome in false negatives because mysterious attacks have not been previously defined. In addition, most systems only analyze and build profiles and patterns after the

fact. These profiles and patterns of activities are subsequently incorporated into rule-based systems to know the future attacks. To be able to detect intrusions as and when it occurs, there is a need for the intrusion detection system to be a real-time system. There is a necessity to automatically build profiling data precise for each customer or class of users that can be used to determine standard actions for a user to lessen the occurrence of false alarms and to get better detection. There is a requirement for a system that can notice suspicious actions, determine the source and institute autonomous responses. There is also a call for the intrusion detection system to take routine action, with no waiting for a human administrator to mediate and act, to lessen the effects of an intrusion and to avoid future actions. There is a huge requirement to coordinate information transfer within host, multi-host and also in network environments so that responses to intrusions can be synchronized. In addition, there is a need to unite the above listed capabilities with real-time monitoring of log audit files, port check revealing capability and session monitoring. Feature selection is the process of identifying and withdrawing as much of the irrelevant and redundant attributes as possible [8]. Feature selection prior to learning can be valuable. Reducing the dimensionality of the data lessens the size of the hypothesis space and allows algorithms to operate faster and more effectively [4]. Particularly, in IDS data mining, feature selection may provide us with the means to reduce the number of network features made while still maintaining accuracy and reducing false negative and false positive rates.

The remaining manuscript is structured as follows: Section II presents the details of the related research work on the problem; Section III gives the description of the Dataset. Section IV presents the Objective and Problem definition. Section V gives proposed method. Section VI presents an overview of the techniques employed. Sections VI I and VIII gives the experimental results and conclusion.

The main focus of the manuscript it to build a system which uses various pre-processing methods such as Feature Selection [9] [13] and Discretization. With the help of Feature selection the appropriate features are selected and due to Discretization the data are descritized which can be applied to various classifier algorithms like a support vector machine, KNN etc.

## RELATED WORK

This document presents a literature review of the few areas that covers a span of research. The data set (NSL_KDD) is freely accessible for researchers through the website.

Datta H. Deshmukh, Tushar Ghorpade, Puja Padiya [1] conducted an experimental analysis on NSL_KDD data sets. They built a model which has more focus on how to increase the accuracy of the classifier. For this reason the project had implemented various pre-processing steps on the existing NSL-KDD dataset like feature selection and discretization. The proposed classifier model tries to overcome the problem of high dimensionality of the dataset by selecting the proper feature selection algorithm such as fast correlation based filter algorithm. The next important issue is to choice of the appropriate algorithm for classifier. The model has implemented with various classifiers such as Hidden Naïve Bayes classifier and NBTree classifier. After implementation the proposed classifier model has improved the accuracy of the classifier and decreased the Error rate.

MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani [2] conducted a experiments on simulator log data (KDD CUP), they concluded that there are significant issues which extremely affects the performance of classifier model, and results in a very underprivileged evaluation of anomaly detection approaches. To address these issues, they have proposed a new data set, i.e. network simulator logs knowledge discovery data (NSL-KDD) [3]. The benefits of having NSL_KDD over the original KDD data set [11] are listed below:

1. The NSl_KDD does not contain redundant records in the train set; hence the classifiers will not be influenced towards more recurrent tuples.
2. No replica of the tuples in the proposed test sets; due to that, the performance of the classifier models are not biased, hence which gives better detection rates on the frequent tuples.
3. The number of tuples in the train and test sets is reasonable, which makes it reasonable to run the experiments on the complete set.

Adetunmbi A. Olusola, Adeola S. Oladele and DaramolaO. Abosede [4] projected the significance of each feature [5] in KDD 99 intrusion detection dataset to the detection of each class. They found that selecting the right

attributes is challenging task, but it must be performed to lessen the number of features for the sake of enhancement in processing speed and to eliminate the irrelevant, redundant and noisy data for the sake of predictive accuracy.

S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas [10] discussed at various parameters which affects the success of Machine Learning (ML) on a given task. They concluded that discretization step will significantly lessen the number of possible values of the continuous feature since a large number of possible attribute values contributes to slow and ineffective practice of inductive machine learning (ML). The crisis of choosing the interval borders and the exact rarity for the discretization of a numerical value series remains an open crisis in numerical feature handling.

Sunitha Beniwal et al. [12] discussed different Classification and Feature Selection technique in data mining. Classification is a process of discovering classes of mysterious data. Feature selection is a technique used to evade over fitting and to enhance the model performance to afford quicker and further cost effective models. They discussed various methods for classification like Bayesian, decision trees, etc. With they concluded that filtering (removing irrelevant features) is the initial step of data mining. Filtering can be done using different feature selection techniques like wrapper, filter, and embedded technique. The data mining responsibilities can be broadly classified in two categories: the descriptive one and the predicted one. In descriptive mining it represents the broad-spectrum properties of the data in the list. In predictive mining it performs the inference on the current data in order to give the results for predictions.

Research [16] shows that the legitimate way of evaluating features is through the error rate of the classifier being designed. The classification error rate is used as a performance pointer for a mining task, for a selected attribute subset; simply conduct the "before-and-after" experiment to compare the error rate of the classifier learned on the full set of features and that learned on the selected subset [17].

## DATASET

The NSL KDD Dataset [3] is one of the few currently available public data sets. The majority of the experiments in the intrusion detection domain is performed on this dataset. Since our model is based on supervised learning methods, NSL KDD is the available dataset that provides labels for both training and test sets. The NSL KDD dataset is the public dataset on network events that contains a comprehensive set of labelled intrusion events. This dataset is quite large in terms of both number of instances and a number of features, and it provides interesting characteristics on the distribution of events and on the dependencies between features. These interesting characteristics and challenges of the dataset make it much more appropriate for use as a benchmark in intrusion detection studies. The dataset contains training data that include 7 weeks of network traffic in the form of Transmission Control Protocol (TCP) dump data consisting of approximately 5 million connection records, each of which is approximately 100 bytes. The test data included 2 weeks of traffic, with roughly 2 million connection records. The 10% NSL KDD99 dataset was used as the training dataset in the competition.

## PROBLEM STATEMENT

This Section is to describe the problem definition associated with IDS data mining. The dataset which is being used for the experiments is NSL KDD which is freely available public dataset in the Intrusion Detection domain here the proposed system helps to Increase the accuracy and decreases the error rate of classifiers, but The problem occurred here is the High dimensionality of the dataset which affects the accuracy of classifiers. So the solution for this problem is given in the proposed system by applying the ideal feature selection algorithm I. e. Variance-Threshold (VT) algorithm which reduces the dimensionality of the dataset in pre-processing part. Here the selection of a proper algorithm for classifier also plays a very important role. For those in the proposed system the SVM, KNN algorithms are used which helps to increase the accuracy of the classifier.

ISSN 2349-4506
Impact Factor: 2.265

Global Journal of Engineering Science and Research Management

## THE PROPOSED METHOD

The proposed method uses NSl_KDD data sets for experiments. In the proposed method the classifier model is built by supervised learning method. The sequences of steps (Fig 1) are as follows

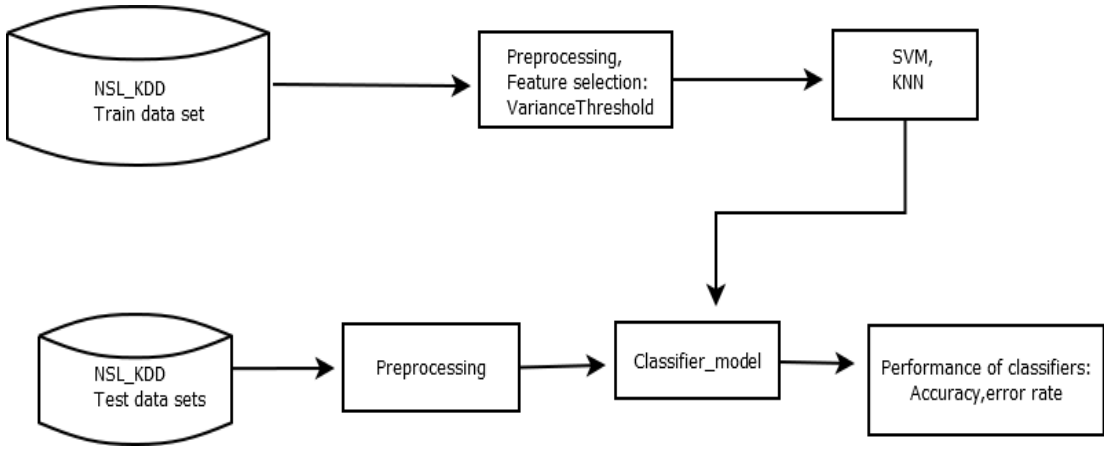


***Fig 1: The proposed classifier model***

Step 1; Segregate the dataset into two distinct, i.e. training set and testing set. Step 2: Perform Pre-processing. It performs the feature selection with the help of Variance-Threshold filter method. Feature selection is an initial step of data pre- processing. This step is quite effective in reducing dimensionality, removing inappropriate data, increasing learning accuracy. This method can identify relevant features for further processing. Step 3: Various Classifiers are implemented like SVM, KNN etc. and finally Accuracy and Error rates are calculated.

## OVERVIEW OF THE TECHNIQUES EMPLOYED

The following techniques are applied to classify the NSL_KDD: Support Vector Machine (SVM), K-Nearest Neighbor.

**KNN (k-Nearest-Neighbor)**

A Nearest Neighbor Classifier based on learning by analogy. The training samples are described by n-dimensional numeric attributes. Each sample describes a point in a n-dimensional space. In this way, all of the training samples are stored in a n-dimensional pattern space. In most of the cases when given an unknown sample, a K-nearest neighbor classifier searches the pattern space for the k closest to the unknown samples. These k training samples are the k nearest neighbors of the unknown samples. Here closeness is defined in terms of Euclidean distance, where the Euclidean distance between two points

$X= (x_1, x_2, x_3....., x_n)$ and $Y= (y_1, y_2, y_3,.....y_n)$ is

$$d(X,Y) = \sqrt{(x1-y1)^2+(x2-y2)^2+\cdots+(xn-yn)^2} \quad (1)$$

The indefinite records are assigned the most common class among its k nearest neighbors. When the given instance k=1, the indefinite sample is allocated the class of the training sample that is closest to it in pattern space. The KNN technique predicts that the entire sampling set includes not only the data in the set, but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the sampling set must be calculated. Only the k nearest entries in the sampling set are considered further. The new item is then classified to the appropriate class that contains the most items from this set of k closest items.

**Support Vector Machine (SVM)**

Support Vector Machines [13] are basically binary classification algorithms. Support Vector Machine (SVM) is a classification scheme derived from the statistical learning concept. It has been applied successfully in fields such as text categorization, handwritten character recognition, image classification, bio sequences analysis, etc. The SVM separates the classes with a decision surface that scale up the margin between the classes. The surface is

often referred as the optimal hyper-plane, and the selected data points nearer to the hyper-plane are said to be a support vectors. The support vectors are the critical components of the training set. The mechanism that describes the mapping procedure is called the kernel function. The SVM can be customized to become a nonlinear classifier through the use of nonlinear kernels. The output of SVM classification is the decisive principles of each pixel for each class, which are used for probability estimates. The probability values symbolize "true" probability in the sense that each probability falls in the range of 0 to 1, and the sum of these parameters for each pixel equals 1. Classification is then done by choosing the highest probability. The SVM is a dominant algorithm based on recent advances in statistical learning theory projected by Vapnik [13]. SVM is a learning system that makes use of a hypothesis space of linear functions in a high dimensional space, trained with a learning algorithm from optimization theory that implements a learning bias discovered from statistical learning theory. The training samples that are very nearer to the maximum margin hyper-plane are called support vectors. All other training samples are irrelevant for defining the binary class boundaries. The support vectors are then used to build an optimal linear separating hyper-plane or a linear regression function (in case of regression) in this feature space.

## RESULTS AND DISCUSSION

This section gives the details of the technical platform required for conducting the experiments. In this work we will be using python 2.7 and all its packages like sklearn. Numpy, Scipy etc. for Implementation of various algorithms defined in the project. We first observed the results obtained from the experiments performed on the NSL KDD Dataset for Intrusion detection problem by using a feature selection method. After pre-processing the resultant features are used to train the machine and then classifier and computed Accuracy and Error Rate as important performance measures. Accuracy is the fraction of correctly classified instances, and error rate is the fractions of misclassified instances in a dataset. These two measures effectively summarize the overall performance of the classifier. Here they hold out method is used to estimate the classifier accuracy. The table 1 shows the obtained values. The fig 2 describes the chart of accuracy and error rates obtained by different classifier algorithm. The accuracy of a model for a given set is the percentage of test samples that are correctly classified by the model.

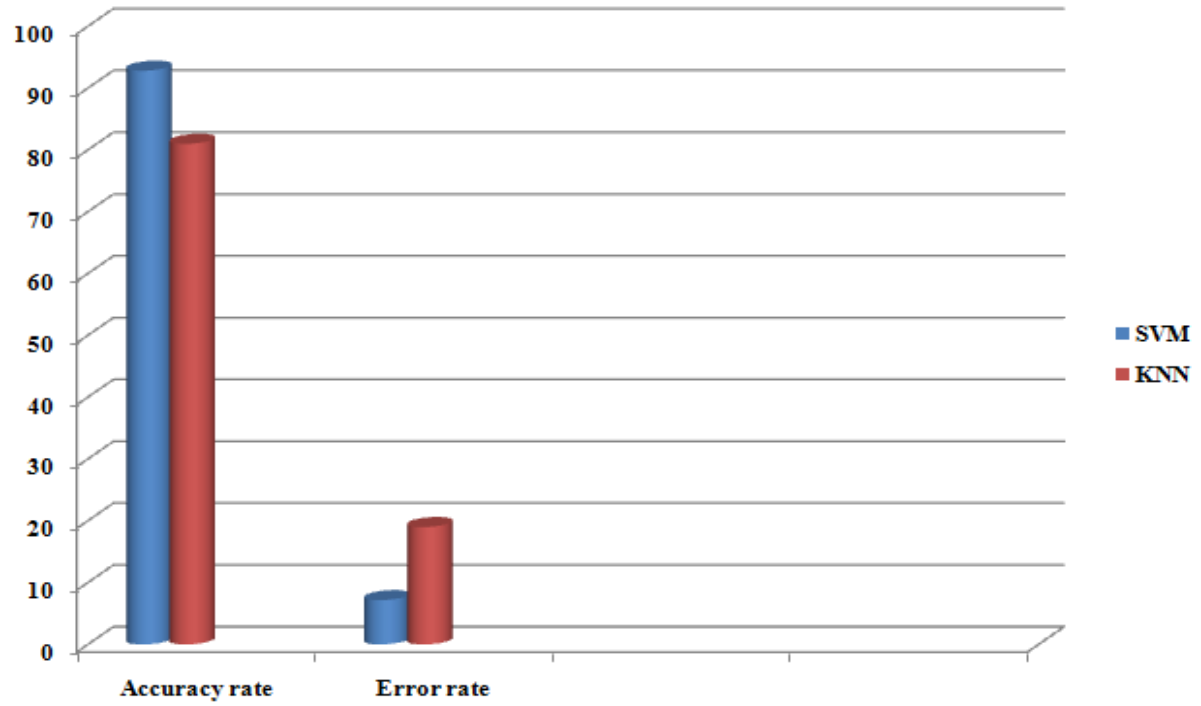


*Fig 2: The Accuracy graph*

***Table 1: The comparison table for results***

| Parameters | Support vector machine | K-Nearest Neighbour |
|---|---|---|
| Accuracy rate | 92.83 | 81 |
| Error rate | 07.17 | 17 |

Obtained Accuracy (SVM) = (Number of samples correctly classified / Total test samples)
= (15204 / 16809)
Obtained Accuracy = 92.83%
Error rate = 1- Obtained Accuracy rate
= 1-0.9283
Error rate = 0.0717 i.e. 07%

## CONCLUSION

The proposed system explains the need to apply data mining methods to network events to classify network attacks and improve the accuracy of the classifier. The System has more focus on how to increase the accuracy of the classifier. The proposed system attempts to overcome the problem of High Dimensionality of the Dataset by selecting the proper feature selection algorithm such as Variance-Threshold. The next important issue is to section of proper algorithms for classifier so in the existing system Naive Bayes has the drawback of conditional independence assumption. To overcome this issue the project has implemented various classifiers such as support vector machine, k-nearest-neighbor classifier. After implementation the proposed system has improved the accuracy of the classifier and decreased the Error rate.

## REFERENCES

1. Datta H.Deshmukh, Tushar Ghorpade, Puja Padiya "Improving Classification Using Preprocessing and Machine Learning Algorithms on NSL-KDD Dataset", 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India
2. Tavallaee, Mahbod, et al. "A detailed analysis of the KDD CUP 99 data set." Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.
3. The NSL KDD Dataset. [Online]. Available http://nsl.cs.unb.ca/NSL-KDD/,On Dated July 30, 2013.
4. Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola O. Abosede. "Analysis of NSL KDD'99 Intrusion Detection Dataset for Selection of Relevance Features." Proceedings of the World Congress on Engineering and Computer Science.Vol. 1. 2010.
5. Kumar, Manish, M. Hanumanthappa, and TV Suresh Kumar. "Intrusion Detection System using decision tree algorithm." Communication Technology (ICCT), 2012 IEEE 14th International Conference on. IEEE, 2012.
6. Kayacik, H. Gnes, A. NurZincir-Heywood, and Malcolm I. Heywood. "Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets." Proceedings of the third annual conference on privacy, security and trust. 2005.
7. Tavallaee, Mahbod, et al. "A detailed analysis of the KDD CUP 99 data set." Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.
8. Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.
9. Shobhan kuma,Naveen D.C "A Survey on Improving Classification Performance Using Data Preprocessing And Machine Learning Methods on NSL-KDD Data" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume – 5 Issue -04 April, 2016 Page No. 16156-16161
10. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas Data Preprocessing for Supervised Leaning IJCS VOLUME 1 NUMBER 2 2006 ISSN 1306-4428
11. KDD-Cup.-(1999), http:// kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,KDD Cup 1999 Data Retrieved July 29, 2013.
12. Beniwal, Sunita, and Jitender Arora. "Classification and feature selection techniques in data mining." International Journal of Engineering Research and Technology. Vol. 1. No. 6 (August-2012). ESRSA Publications, 2012.
13. V.N. Vapnik, "Statistical Learning Theory", John Wiley,New York, 1998.

14. Nguyen, HuyAnh, and Deokjai Choi. "Application of Data Mining to Network Intrusion Detection:classifier selection model," Challenges for Next Generation Network Operations and Service Management. Springer Berlin Heidelberg, 2008. 399-408
15. Depren, O., Topallar, M., Anarim, E.,and Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert systems with Applications, 29(4), 713-722.
16. Sarojini Balakrishnan et.al "SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases", *2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008)*
17. H. Liu and H. Motoda "Feature Selection for Knowledge Discovery and Data Mining", Boston: Kluwer Academic Publishers, 1998.
18. Yu, Lei, and Huan Liu. "Feature selection for high dimensional data: A fast correlation based filter solution." In ICML, vol. 3, pp. 856-863. 2003.
19. Zhang, Harry, Liangxiao Jiang, and Jiang Su. "Hidden naive bayes." Proceedings of the National Conference on Artificial Intelligence. Vol. 20. No. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005. 20 .C.W. Hsu, C.C. Chang and C.J. Lin, "A practical guide to support vector classification", http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide .pdf, 2003
20. Kulkarni, A., & Bush, S. (2006). Detecting distributed denial-of-service attacks using Kolmogorov complexity metrics. Journal of Network and Systems Management, (1), 69–80.