IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. III (Sep – Oct. 2014), PP 57-61
www.iosrjournals.org

Using Ensemble Methods for Improving Classification of the KDD CUP '99 Data Set

Rohan D.Kulkarni
(Department of Computing Sciences/VIT University, India)

Abstract: *The KDD CUP '99 data set has been widely used for intrusion detection and pattern mining in the last decade or so. Umpteen number of experiments pertaining to classification have been conducted on it.Many researchers have dedicated their resources for analysing this data set. But it has yet to be analysed by using Ensemble methods of classification. This paper contains experimental results obtained after classifying 10 % of the KDD CUP '99 data set using ensemble methods like Bagging,Boosting and compares their performance with the standard J-48 classification algorithm.Weka experimenter has been used to classify the 494020 records using the aforementioned classifiers and the advantages of ensembling have been discussed in accordance with the obtained results..*

Keywords: *Bagging, Boosting, Data mining, Ensemble classifiers, KDD CUP '99 data set, Weka*

I. Introduction

The term Data mining is used to describe the process of knowledge discovery from data. It is a process that involves analysis and summarization of a huge amount of data stored in a warehouse and extraction of non-obvious and intricate patterns. These patterns can be used further to support decision making in industries in order to reduce costs, increase revenues or both. Medical industries use this valuable information to gauge the needs of the people and enhance their services. The scope of Knowledge Discovery from Data (KDD) is not limited to any specific industry. Data scientists have been analyzing historical data from organizations for years, but the recent advancement of computational power, increase in disk storage space and emergence of user friendly tools for mining have made the task much simpler.

Classification and Clustering are the two major ways of summarizing data in hand. While Classification is used to aggregate labelled data, clustering is used to summarize data without any pre-defined labels. Classification is the process of defining a model that identifies intricate differences between data points so as to be able to classify objects whose labels are unknown. The end result of both classification and clustering are groups of similar data objects which can be used for further study. As the KDD dataset used for analysis in this paper is a labelled dataset, classification has been used as the method of aggregation.

A very important application of data mining is its use for enhancing security of computational systems.For applying these techniques in the domain of security, we need to have a dataset upon which to deploy the machine learning algorithms.So the KDD CUP '99 dataset, a very commonly used dataset in the field of information security research has been utilized here.The network intrusion dataset contains 41 attributes and is a 10% subset containing about 500,000 tuples.

1.1 KDD CUP '99 Dataset Description

The KDD '99 training dataset is a collection of about 494,020 records which account for 10% of the complete dataset. Every tuple is a single connected vector described by 41 attribute values and exactly one label of either 'normal' or an 'attack'. Every record is checked for abnormal behavior and if found to deviate from the convention, is considered as an attack. Attacks labelled as 'normal' are records with normal behavior whereas all records which are not 'normal' are considered as malicious attacks. Every attack is classified into exactly one of the following categories:

1. Denial of Service Attack (DOS): This is a type of attack where the attacker makes an attempt to make a resource unavailable for its intended users. He makes the resources preoccupied so as to deny legitimate users an opportunity to use that particular resource.DOS attacks are very common in large banking and booking systems where unavailability of a resource can lead to serious financial losses.
2. User to Root Attack (U2R): This is one of the most dreaded cyber-attacks in the industry. Here the attacker gains access to an end user's account and tries to obtain root access to the entire information system. Having gained access, the attacker can then disrupt the normal functioning of the entire system. Examples include 'buffer_overflow','loadmodule etc.
3. Remote to Local Attack (R2L):The main intention of this type of attack is to gain local access to an end user machine.The attacker sends a packet to a machine over a network and then tries to gain access to

an account on that machine.R2L includes attacks like 'warezclient','multihop','ftp_write','imap','guess_passwd' etc.

4. Probing Attack (PROBE): The attacker performs the Probe Attack in order to obtain more information on the computers in a particular network. This type of attack is extrinsically benign but valuable information gained from it can lead to harmful attacks in the future.'portsweep','satan','nmap','ipsweep' are some of the probing attacks.

The KDD dataset consists of 494,020 records out of which 97,277(19.69%) are 'normal' attacks while 391,458 (79.24%) are DOS, 4107(0.83%) are PROBE.R2L attacks comprise of 1,126 (0.23%) of the data and U2L attacks make up about 0.01% of the dataset in consideration.

II. Related work

KDD CUP '99 dataset is one of the most widely studied datasets from the UCI Data repository .Machine learning algorithms have been used to dig deep for emergent patterns .Most of these algorithms are based on clustering or classification .One of the simplest but very efficient clustering algorithm i.e. k-means has been used on this data set[1].This paper suggests the division of the entire dataset into 1000 clusters and uses the Oracle Data mining Module(ODM) to identify the main attacks occurring on a particular protocol. Interesting patterns suggesting the vulnerability of the TCP/IP protocol towards attacks emerged in their study.

Detection of anomalies and outliers in data is another interesting study. However there are some problems in the current approaches being utilized for anomaly detection. The above problem has been solved by suggesting an entirely new data set –NSL KDD which is a subset of the KDD CUP data set and does not suffer from the shortcomings of the earlier dataset[2].There have also been experiments to determine whether a particular kind of machine learning algorithm is better at clustering a particular type of attack in the KDD CUP '99 data set[3].The study showed that the application of specific algorithms increased the accuracy of results for particular attack types while clustering and notable enhancements were seen in DOS, Probing and U2R attack types.

Ensemble classification method is the use of multiple classifiers one after another on a data set so as to improve the accuracy of the final results. The problem of Text Classification (TC) involves usage of Artificial Neural Networks and Machine Learning algorithms for achieving results. Recent studies involving using an ensemble of classifiers for TC have achieved promising results [4].'Lior Rokach' in his paper titled 'Ensemble based classifiers' [5] provides a comprehensive review of the most important ensemble algorithms applied on data. Research on Support Vector Machine (SVM) ensemble classifiers is gaining popularity now-a-days and the authors of [6] provide an empirical study on the same by using these methods on some datasets from the UCI repository.

So, even though it is a widely accepted fact that ensembles improve clustering and classification results when applied on data, none of the researchers have tested it on the KDD CUP '99 dataset. The main objective of this paper is to apply the most popular ensemble classifiers on this data set and compare their results with the standard J-48 algorithm and report the empirical results.

III. Ensemble classifiers

In statistics and machine learning, ensemble classifiers are used to improve predictive results obtained by using its constituent algorithms. An ensemble combines a number of 'weak' learners and aggregates their results into one 'strong' learner. An ensemble is a 'supervised' learning algorithm itself. This is because it can itself be trained on labelled data and later used to predict labels for previously untested data. Ensembles thus work in a two-step process. Firstly, a set of classifier methods are learnt.And later the results of these methods are combined to obtain higher accuracy. Ensembling many algorithms has many advantages over using a singular classification method. Most important advantage is that the results of the classification are less dependent on the peculiarities of a single algorithm. So basically, ensembling ensures a more generic inspection of data at hand. Another advantage is that ensembling of different methods makes the entire model more expressive and unbiased than a single model. Actually, it has been proven experimentally that typically, ensembles tend to produce better results when there is a significant diversity among the algorithms being combined.

There are a number of commonly used and implemented ensemble classifiers. Some of them being Bayes optimal classifier, Bootstrap aggregating, Boosting, Bayesian model averaging, Bayesian model combination, Stacking and voting. While all of them are known to produce significantly improved results, this paper involves the usage of Bagging and Boosting algorithms on the KDD data set. They are explained below:

3.1: Bootstrap aggregating (bagging)

Bagging is a machine learning ensemble algorithm used to improve the accuracy of the algorithms used in statistical classification and regression experiments .Most importantly, it reduces variance and resolves the problem of over fitting the data. The algorithm for bagging is as follows:

Algorithm

1. for m = 1 to M // M ... number of iterations
    a) Draw (with replacement) a bootstrap sample S_m of the data
    b) Learn a classifier C_m from S_m
2. for each test example
    a) Try all classifiers C_m
    b) Predict the class that receives the highest number of votes.

Given any dataset D containing 'n' tuples, the bagging algorithm creates 'm' different datasets from the given dataset each containing n' tuples with replacement. Such a sample is known as a 'bootstrap' sample. Every model is fitted with a bootstrap sample and combined by taking a mean of the outputs.

3.2: Boosting

The main aim of the boosting algorithm used for supervised learning is to reduce the bias in the different classification techniques used. A weak learner is an algorithm with better accuracy than a random surmise. Boosting combines a number of such weak learners to form a strong supervised learning algorithm. The algorithm for boosting is given below:

Algorithm

1. Initialize example weights $w_i = 1/N$ (i = 1...N)

2. for m = 1 to M // M ... number of iterations
    a) Learn a classifier C_m using the current example weights
    b) Compute a weighted error estimate
        $err_m = \Sigma wi$ of all incorrectly classified e_i/Sum of weights
    c) Compute a classifier weight
        $a_m = 0.5 \log ((1-err_m)/err_m)$
    d) for all correctly classified examples e_i: $w_i = w_i e^{-am}$
    e) for all incorrectly classified examples e_i : $w_i = w_i e^{am}$
    f) Normalize the weights w_i so that they sum to 1

3. for each test example:
    a) Try all classifiers C_m
    b) Predict the class that receives the highest sum of weights a_m.

Weak learners are stacked on top of one another iteratively. Once a new weak algorithm is added, the data is reweighted. The tuples that are misplaced gain importance over the correctly classified ones. Thus the next classifier added concentrates on the classification of the misplaced data. Hence eventually most of the data gets classified with great accuracy. In this experiment, the data has been subjected to the AdaBoost algorithm which is an adaptive boosting technique which is a very popularly used boosting technique in machine learning.

IV. Experimental results

The experiment on the KDD CUP '99 intrusion detection data set was performed using the Weka Experimenter version 3.7.4.The Weka experimenter is an Open Source very easy to use software for performing Knowledge Discovery from Data. Many generally used algorithms as well as complex stacking and voting algorithms have been implemented in Weka for better understanding and summarization of the data set at hand. The experiment in this paper was performed on an Intel 1st generation core i7 processor with a clock speed of 2.20 GHz and 4 cores. A total memory space of 2 GB was allocated for the experiment. The general algorithm used for comparison here is the J-48 decision tree making algorithm implemented in Weka. The results of the experiment have been delineated below:

4.1: Ranks of the three algorithms

The analysis shows that the AdaBoostM1 algorithm is better than both the other algorithms used in the experiment. In the figure below, it is clear that the AdaBoost algorithm is better than 2 and inferior to none of the algorithms used.

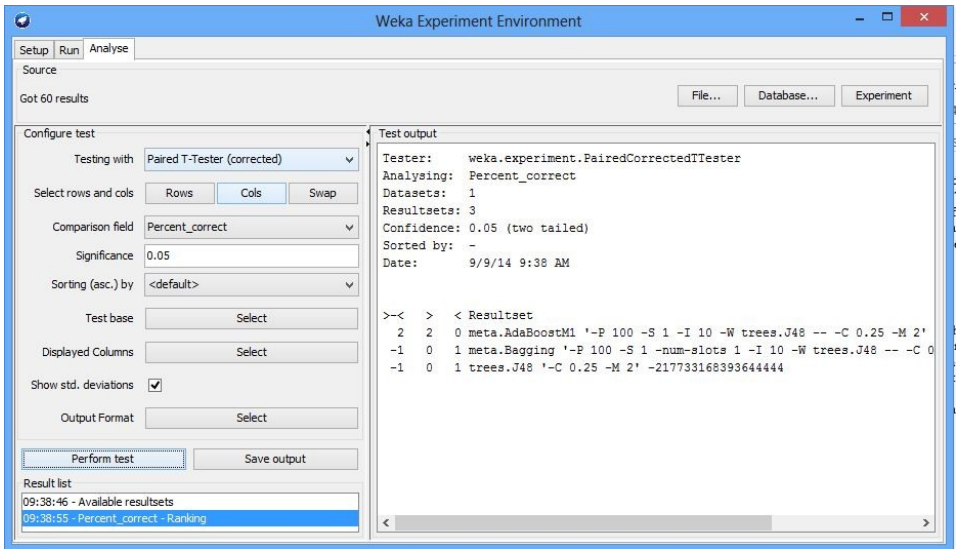

Fig 1.Ranks of the algorithms used in the experiment.

4.2: Incorrect classification

Form the charts given below it can be inferred that of the three methods, AdaBoost made the least percentage and number of incorrect classifications of objects in the given data.

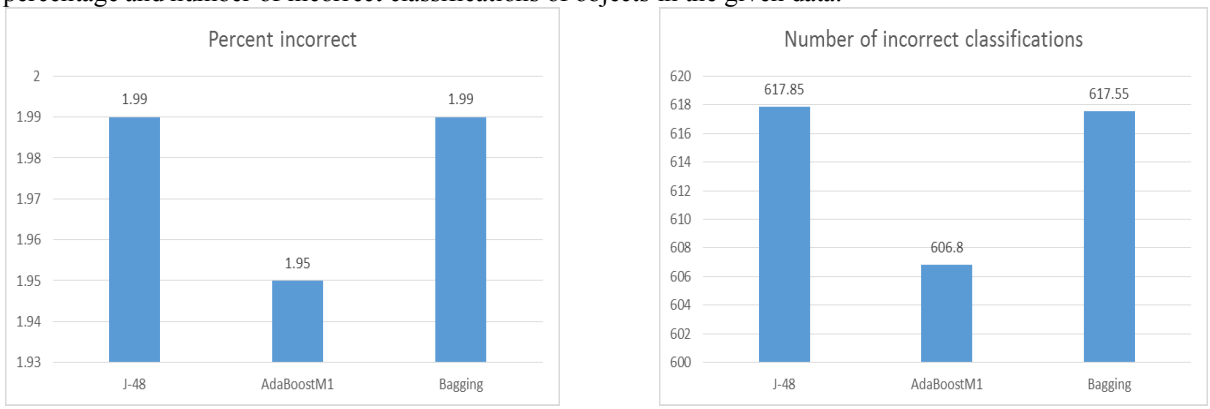

Fig 2.Percentage and number of incorrect classification done by the 3 algorithms

4.3:Error rate

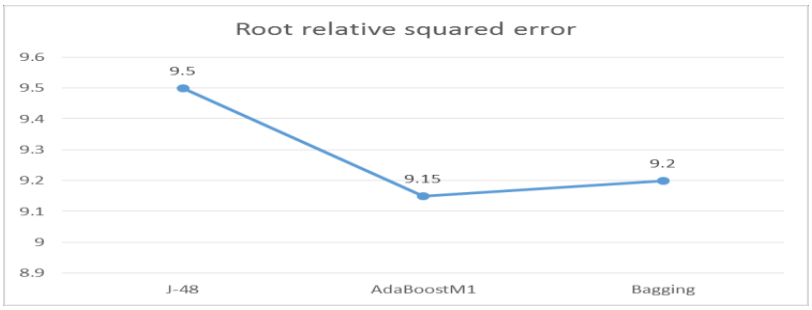

Fig 3.Root relative squared error

4.4: Number of False positives and False Negatives

False positives are those values in the data that are actually negative but are falsely classified as positive. Here positive means that the value belongs to a cluster and negative means that it does not belong to a

particular cluster.From the bar graphs below,we can see that while the AdaBoostM1 algorithm is better than the other two as far as false positives are concerned,bagging is the best to avoid false negatives.

Fig 4.False positives and false negatives

V. Conclusion

The experimental results clearly indicate that among the general J-48, Bagging and AdaBoostM1 algorithm, the latter proves to be the least error prone and efficient when employed on the KDD CUP '99 Intrusion detection dataset. It performs better as its error rate is significantly lower and the probability of incorrect classification is also minute. The main contribution of this paper are the results that prove that the Ensemble method of Boosting is indeed better for classifying the Intrusion detection dataset.

However it is of paramount importance that this data set is further studied using other ensemble methods so as to find out whether there is an algorithm or a combination of them (ensemble) that is better than the AdaBoostM1 algorithm.

References

[1] Mohammad Khubeb and Shams Naahid, Analysis of KDD CUP '99 Dataset using Clustering based Mining, International Journal of Database Theory and Application, 6(5), 2013, 23-34.

[2] Mahbod Tavallaee,Ebrahim Bagheri,Wei Lu and Ali .A.Ghorbani, A Detailed Analysis of the KDD CUP '99 Dataset, Proc.of 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications,978-1-4244-3764-1/09.

[3] Maheshkumar Sabhnani and Gursel Serpen, Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, Proc .Of the International Conference on Machine Learning;models,Technologies and Applications, Las Vegas,Nevada,USA 1-932415-11-4.

[4] Yan-Shi Dong and Ke-Song Han, Boosting SVM Classifiers by Ensemble, Proc.of 2005 ACM WWW Conference,Chiba,Japan 1-59593-051-5/05/0005.

[5] Lior Rokach, Ensemble based Classifiers, Springer Science+Business Media, 2009.

[6] David Opitz and Richard Maclin, Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, 11, 1999, 168-198.