International Journal of Computer Science and Engineering (IJCSE)
Vol.1, Issue 1 Aug 2012 81-88

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

COMPARATIVE ANALYSIS OF TWO ALGORITHMS FOR INTRUSION ATTACK CLASSIFICATION USING KDD CUP DATASET

N.S.CHANDOLIKAR[1] & V.D.NANDAVADEKAR[2]

[1]Reader, Vishwakarma Institute of Technology, Pune, India

[2]Director –MCA, Sinhgad institute of management, Pune, India

ABSTRACT

In today's interconnected world , one of pervasive issue is how to protect system from intrusion based security attacks. Significance of Intrusion detection system (IDS) in computer network security well proven . Mining approach can play very important role in developing intrusion detection system. Classification is identified as an important technique of data mining. This paper evaluate performance to two well known classification algorithms for attack classification. Bayes net and J48 algorithm are analyzed The key ideas are to use data mining techniques efficiently for intrusion attack classification.

KEYWORDS: Intrusion Detection System, Bayes Net, J48 Classification Algorithm.

INTRODUCTION

The security of a computer system is compromised when an intrusion takes place. An intrusion can be defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource". Intrusion Detection[1] is the unrelenting active attempts in discovering or detecting the presence of intrusive activities

Data Mining

Data mining [2] [3][4] is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Data mining can be used for solving the problem of network intrusion based security attack. It has Ability to process large amount of data and reduce data and by extracting specific data, With this Easy data summarization and visualization that help the security analysis.

Intrusion Detection System

Intrusion Detection System (IDS) can detect, prevent and more than that IDS react to the attack. Therefore, the main objective of IDS is to at first detect all intrusions at first effectively. This leads to the use of an intelligence technique known as data mining/machine learning. These techniques are used as an alternative to expensive and strenuous human input. •IDS can provide guidelines that assist you in the vital step of establishing a security policy for your computing assets.

Detection method in IDS [5][6][7] can be divided into two categories: anomaly detection and misuse detection categories.

Signature-Based IDS

Network traffic is examined for preconfigured and predetermined attack patterns known as signatures. It is widely available, it uses known patterns as it is easy to implement but they cannot detect attacks for which it has no signature and they are also prone to false positives since they are commonly based on regular expressions and string matching. Since they are based on pattern match, signatures usually don't work that great against attacks with self-modifying behavior.

Anomaly-Based IDS

Anomaly-based IDS works on a performance baseline based on normal network traffic evaluations. It sample current network traffic activity to this baseline in order to detect whether or not it is within baseline parameters. Data mining techniques can be used for intrusion detection efficiently.

INTRUSION DETECTION DATASETS

KDDCup'99 Data Set

The data set used to perform the experiment is taken from KDD Cup '99[8][9], which is widely accepted as a benchmark dataset and referred by many researchers. "10% of KDD Cup'99" from KDD Cup '99 data set was chosen to evaluate rules and testing data sets to detect intrusion. The entire KDD Cup '99 data set contains 41 features. Connections are labeled as normal or attacks fall into 4 main categories.

1. DOS:- Denial Of Service
2. Probe:- e.g. port scanning
3. U2R:- unauthorized access to root privileges,
4. R2L :- unauthorized remote login to machine.

In this dataset there are 3 groups of features: Basic, content based, time based features.

- Training set consists 5 million connections.
- 10% training set - 494,021 connections
- Test set have - 311,029 connections
- Test data has attack types that are not present in the training data .Problem is more realistic
- Train set contains 22 attack types
- Test data contains additional 17 new attack types that belong to one of four main categories.

PROPOSED SYSTEM DESCRIPTION

For the experiment two well performed algorithms are used first is bayes based bayes net and second is decision tree J48 .

The algorithms used in this investigation are briefly described

in the following paragraphs

Classification

Classification [1] [2] data mining technique Classification maps a data item into one of several pre-defined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class. there are many types of classifiers are available like tree, bayes, function ,rule . basic aim of classifier is predict the appropriate class.

Decision Tree

Decision tree [1] [2] [10] [11] [12]is an important method for data mining, which is mainly used for model classification and prediction. This predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

J48 Algorithm

The J48 is a Decision tree classifier algorithm. In this algorithm for classification of new item, it first needs to create a decision tree based on the attribute values of the available training data. It discriminate the various instances and identify the attribute for the same. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

Bayes Net

Bayes net[13] are based on bayes theorem. bayes net is an directed acyclic graph. for the formation of bayes net conditional probability is used. This algorithm assumes that there are no missing values and all attributes are nominal.

Feature selection [14] is one of the common terms used in data mining. It is used to reduce inputs to a manageable size for processing and analysis. Many tools and techniques are available for the same. Feature selection is used for imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, and also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis.

Feature selection for intrusion detection is an important factor for the success of intrusion detection system. supervised discrete filter is used for attribute selection.

EXPERIMENTAL SETUP

To assess the effectiveness of the algorithms for proposed intrusion detection, the series of experiments were performed in Weka. The java heap size was set to 1024 MB for weka-3-6.

KDD 99 dataset is investigated to identify the relevance of each feature in intrusion detection.

To test and evaluate the algorithms we use 10-fold cross validation. In this process the data set is divided into 10 subsets. Each time, one of the 10 subsets is used as the test set and the other k-1 subsets form the training set. Performance statistics are calculated across all 10 trials. This provides a good indication of how well the classifier will perform on unseen data.

We used the J48,bayes net algorithm available on the Weka collection of machine learning algorithms. J48 is the Weka implementation of the decision tree learner C4.5. these algorithms are chosen for several reasons: these are well-known classification algorithms. It can originate easily understandable rules and are designed to classify into predefined discrete categories (classes). .

Weka

Weka[13][15] is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface, Java interface.

Performance Measurement Terms

Correctly Classified Instance

The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified.

The percentage of correctly classified instances is often called accuracy or sample accuracy.

Kappa Statistics

Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that your classifier is doing better than chance (it really should be!).

Mean Absolute Error, Root Mean Squared Error, Relative_Absolute_Error

The error rates are used for numeric prediction rather than classification. In numeric prediction, predictions aren't just right or wrong, the error has a magnitude, and these measures reflect that.

Detection of attack is measured by following metrics:

- False positive (FP): Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.
- False negative (FN): Corresponds to the number of detected normal instances but it is actually attacks, in other words these attacks are the target of intrusion detection systems.
- True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.
- True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.
- The accuracy of an intrusion detection system is measured regarding to detection rate and false alarm rate.

RESULTS AND DISCUSSIONS

Our ultimate goal is to evaluate performance of two algorithm for attack classification

Algorithm are evaluated on the bases of true positive(TP) and false positive (FP) rate. Experiment is performed to detect 5 different classes of attacks from the dataset including Dos, U2R, Probe, U2L and normal. The distribution of an attack and normal records are 80%-20%. .based on the experiment association of any feature with attack class is analyzed.

The detection algorithm maps incoming events to attacks and normal activity. The resulting classification can be used to determine the effectiveness of an IDS. Effectiveness is the ability of an IDS to maximize the detection rate while minimizing the false alarm rate (false positive rate). In other words, good IDS reports intrusions when they occur, and does not report intrusions when they do not occur

Table: 1.True Positive and False Negative Rate in J48 and Bayes Net

Sr.No.	Type of attack		Bayes Net	J48
1	Normal	TP	0.969	0.998
		FP	0.003	0.003
2	DOS	TP	0.9915	0.9938
		FP	0	0
3	U2R	TP	0.6	0.2
		FP	0.001	0
4	R2L	TP	0.85	0.75
		FP	0.001	0
5	Probe	TP	0.857	0.988
		FP	0.004	0

For accuracy measurement ,Table II shows the Performance of j48 and bayes net algorithm based on Correctly classified instance, Incorrectly classified instance, Kappa statistics, Mean absolute error, Root mean squared error.

Table 2 Performance of J48 Algorithm

Sr.no.	Parameter	Performance of bayes net in percentage	Performance of J48 in percentage
1	Correctly classified instance	97.3732	99.742
2	Incorrectly classified instance	2.6268	0.258
3	Kappa statistics	0.9571	0.9957
4	Mean absolute error	0.0024	0.0003
5	Root mean squared error	0.0434	0.0145
6	Relative_absolute_error	4.5617	0.656

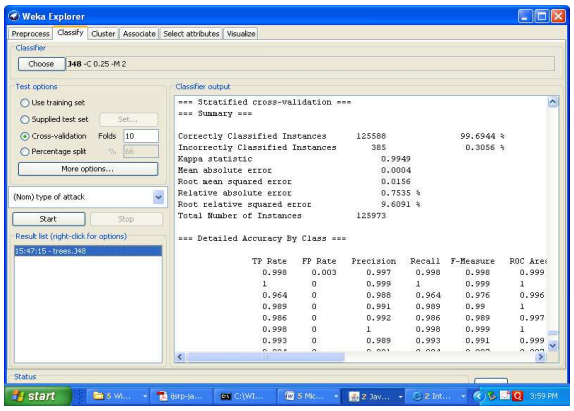

Fig 1 J48 performance Shows The Weka Classification J48 Performance of our Experiment.

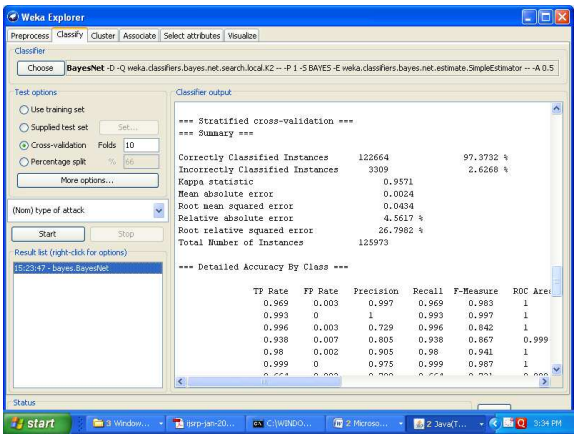

Fig 2 Bayes Net Performance the Weka Classification Bayes Net Performance of Our Experiment.

Based on experiment we can say that normal, neptune and smurf classes are highly related to certain features that make their classification easier. Since these three classes make up 98% of the training data, it is very easy for a Intrusion detection system to achieve good results. There are few features which are not relevant in terms of intrusion detection and there are some which are highly relevant.

Our ultimate goal is to show the impact of choosing algorithm on the performance of IDS. First we identify significantly reduced feature sets using discritization. we compare the relative computational performance of both the tested data mining algorithm.

We find that the feature reduction techniques are able to greatly reduce the feature space .We find that better differentiation of algorithms can be obtained by examining computational performance Bayes Net have faster build times, while J48 is slower. We therefore focus on the build time and classification speed of the algorithms when using each of the feature sets.

Computational performance is particularly important when considering real-time classification of potentially thousands of simultaneous networks flows The both the algorithms achieve greater than 95% accuracy .These preliminary results show that J48 shows better performance accuracy. The J48 algorithm is significantly faster in terms of classification speed and appears to be the best suited for real-time classification tasks.

CONCLUSIONS

Data mining can improve intrusion based security attacks detection system by adding a new level of surveillance to detection of network data indifferences. J48 learning algorithm was found to be performing better than bayes net in terms of better accuracy and lower error rate. Experiment performed on KDD cup dataset demonstrate that J48 algorithm is an efficient algorithm of classification .. Accuracy demonstrated helps to improve efficiency of intrusion detection system.

REFERENCES

1. Stephen Northcutt , Judy Novak “Network Intrusion Detection”, Third Edition, New Riders Publishing
2. Jiawei Han And Micheline Kamber “Data mining concepts and techniques” Morgan Kaufmann publishers .an imprint of Elsevier .ISBN 978-1-55860-901-3. Indian reprint ISBN 978-81-312-
3. 0535-8 .
4. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second edition, 2005. Morgan Kaufmann.
5. T. Lappas and K. P. , “Data Mining Techniques for (Network) Intrusion Detection System," January 2007.

6. L. Zenghui, L. Yingxu, “A Data Mining Framework for Building Intrusion Detection Models Based on IPv6,” Proceedings of the 3rd International Conference and Workshops on Advances in Information Security and Assurance. Seoul, Korea, Springer- Verlag, 2009.

7. Kayacik, G. H., Zincir-Heywood, A. N., “Analysis of Three Intrusion Detection System Benchmark Datasets Using Machine Learning Algorithms", Proceedings of the IEEE ISI 2005 Atlanta, USA, May 2005.

8. Ozgur Depren, Murat Topallar, Emin Anarim, M. Kemal Ciliz. “An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks”. Expert Systems with Applications 29 (2005) 713–722Expert Systems with Applications 29 (2005)713722.www.elsevier.com/locate/eswa.

9. H. Güneş Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood,. “ Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99Intrusion Detection Datasets". Dalhousie University, Faculty of Computer Science, http://www.cs. dal. ca /projectx/

10. The KDD Archive. KDD99 cup dataset, 1999.

11. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

12. Jeff Markey “ Using Decision Tree Analysis for Intrusion Detection: A How-To Guide”. Global Information Assurance Certification Paper , Copyright SANS Institute 2011.

13. Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman ,“Attacks Classification in Adaptive Intrusion Detection using Decision Tree”, World Academy of Science, Engineering and Technology 63 2010

14. E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu,” A Study of Intrusion Detection in Data Mining “,Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011, London, U.K. ISBN: 978-988-19251-5-2 ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

15. Remco R. Bouckaert, Bayesian Network Classi_ers in Weka, September 1, 2004

16. Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede . “Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features. Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA.

17. Muamer N. Mohammada,*, Norrozila Sulaimana, Osama Abdulkarim Muhsinb “A Novel Intrusion Detection System by using Intelligent Data

18. Mining in Weka Environment”, Procedia Computer Science www.elsevier.com/locate/procedia WCIT 2010