Efficient Algorithm for Intrusion Attack Classification by Analyzing KDD Cup 99

Prof. Mrs. N. S. Chandolikar
Dept. of Computer Engg
Vishwakarma Institute of Technology, Pune
India

Dr. V. D. Nandavadekar
Director –MCA
Sinhgad Institute of Management, Pune
India

Abstract- Importance of intrusion detection system (IDS) for network security management is widely accepted. Efficiency of IDS is mainly affected by algorithms used for feature identification and classification. Data mining can be very fruitful for feature selection and intrusion detection. In this paper, we have presented J48 classification algorithm for intrusion detection. To evaluate the performance of the algorithm correctly classified instances, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root relative squared error and kappa statistics measures are applied.

Keywords- Intrusion detection system (IDS), data mining, classification. J48 algorithm.

I. INTRODUCTION

With the increased growth of networked systems and applications, the demand for network security is high. To identify intruion attack ,efficient IDS plays most important role . Intrusion Detection System can detect, prevent and more than that IDS react to the attack. use of an intelligence technique known as data mining/machine learning for intrusion attack detection and classification is very promising. This These techniques are used as an alternative to expensive and strenuous human input.

A. Data Mining:

Data mining [1] [2] is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Data mining [3] [4] 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.

Identification of feature selection in IDS can assist in performance improvement of IDS. Classification models are simple and effective. Predictions of defects from such models can be used to achieve high reliability.

II. OVERVIEW OF INTRUSION DETECTION SYSTEMS

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

A. 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.

B. 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. The system may detect unknown attacks. It may allow one to detect more complex attacks, such as those that occur over extended periods.

Anomaly Detection IDS are attractive conceptually, but they require training and need to classify "normal" traffic.

III. INTRUSION DETECTION DATASETS

A. KDDCup'99 Data Set

The data set used to perform the experiment is taken from KDD Cup '99 [8] [9] [10], which is widely accepted as a benchmark dataset .The data set was chosen to evaluate rules and 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. KDD dataset details:

- Training set - 5 million connections.
- 10% training set - 494,021 connections
- Test set - 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.

IV. FEATURE SELECTION

Feature selection [11] [12] [13] [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.

978-1-4673-1989-8/12/$31.00 ©2012 IEEE

V. CLASSIFICATION METHODS

Classification [1] [2] data mining technique predicts categorical class labels. Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations. New data is classified based on the training set.

A. Decision tree

Decision tree [1] [2] 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.

1) 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.

2) REPTree

REPTree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances.

3) Random Tree

A random tree is a tree drawn at random from a set of possible trees. In this each tree in the set of trees has an equal chance of being sampled. Another way of saying this is that the distribution of trees is "uniform". Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models.

B. Other classification algorithm

1) Rule Induction

Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In rule induction systems the rule itself is of a simple form of "if this and this and this then this". Rules are mutually exclusive and exhaustive.

One Rule algorithm of rule induction adopts a greedy depth-first strategy. Each time it is faced with adding a new attribute test (conjunct) to the current rule, it picks the one that most improves the rule quality, based on the training samples. OneR algorithm steps as are follows:

- Learn one rule at a time, sequentially.
- After a rule is learned, the training examples covered by the rule are removed.
- Only the remaining data are used to find subsequent rules.

The process repeats until some stopping criteria are met

2) Bayes net:

Bayes net 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.

VI. PROPOSED SYSTEM DESCRIPTION

Appropriate classification of intrusion is most important factor for the success of intrusion detection system. KDD 99 dataset is investigated to classify intrusion attacks using data mining algorithms. 10 folds Cross-validation is used while experimentation. Cross-validation is a technique for estimating the performance of a predictive model.

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

We have used algorithms available on the Weka collection of machine learning algorithms. Various categories of algorithms were analyzed. J48 algorithm is used and Performance is compared with two tree based algorithms and two other than tree based algorithm for classification.

J48 is the Weka implementation of the decision tree learner C4.5 [15] . C4.5 was chosen for several reasons: it is a well-known classification algorithm. It can originate easily understandable rules. C4.5 is designed to classify into predefined discrete categories (classes). Since C4.5 only processes features one by one, it does not matter that there are interactions between the features we are using, C4.5 only allows interactions in the form of multi-part conditions. That may result in missing some positive effects of the interaction of features, but does not risk a false positive result, originated from a multivariate test based on false assumptions.

Fig 1 shows the steps performed in this study, which describes the process.

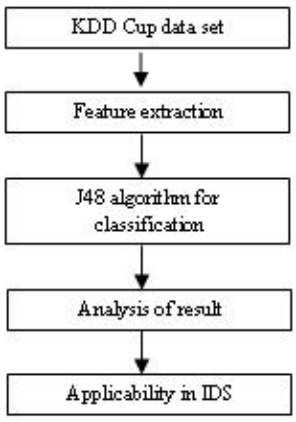

Fig 1
Process steps

Initially KDD cup 99 dataset is taken for experiment on which filter attribute selection is applied for feature selection. After selection of feature j48 classification algorithm is applied to get result.

This study was performed to

- Investigate and pre-process the features in the database and assessing the correctness of the data.
- Define the class attributes which divide the set of instances into the appropriate classes.
- Examine various classification methods for applicability for intrusion attack.
- Make decision on a testing method to estimate the performance of the algorithm.
- Suggest performance improvement in IDS.

A. Weka

Weka[16] 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.

B. Performance measurement terms

1) 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.

2) 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!).

3) 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.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,
- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

Table I shows the confusion matrix for a two class classifier.

TABLE I

CONFUSION MATRIX

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using (1):

$$AC = \frac{a+d}{a+b+c+d} \quad (1)$$

- The *true positive rat*e (*TP*) is the proportion of positive cases that were correctly identified, It is determined using the (2):

$$TP = \frac{d}{c+d} \quad (2)$$

- The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, It is determined using (3):

$$FP = \frac{b}{a+b} \quad (3)$$

- The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly. It is determined using (4):

$$TN = \frac{a}{a+b} \quad (4)$$

- The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, It is determined using (5):

$$FN = \frac{c}{c+d} \quad (5)$$

- The *precision* (*P*) is the proportion of the predicted positive cases that were correct, It is determined using (6):

$$P = \frac{d}{b+d} \quad (6)$$

VII. THE EXPERIMENTAL RESULTS

Proposed intrusion detection approaches are implemented 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.

At first attribute selection process applied using supervised filter attribute selection. To get result at first three different category of classification algorithms are compared. OneR from rule induction category and bayes net from bayes are taken for performance comparison.

- Classification by Bayesian Belief Networks
- Classification by rule induction
- Classification by decision tree induction

Algorithm performance is compared based on correctly classified instance, incorrectly classified instance, Kappa statistics, Mean absolute error, Root mean squared error.

Table II shows the Performance comparison of j48 algorithm with other classification algorithm.

TABLE II

PERFORMANCE COMPARISON OF J48 ALGORITHM WITH OTHER ALGORITHMS

Sr. no.	**Parameter**	**J48**	**OneR**	**Bayes Net**
1	Correctly classified instance	99.742	90.9139	97.3732
2	Incorrectly classified instance	0.258	9.0861	2.6268
3	Kappa statistics	0.9957	0.8478	0.9571
4	Mean absolute error	0.0003	0.0079	0.0024
5	Root mean squared error	0.0145	0.0889	0.0434
6	Relative_absolute_error	0.656	15.0351	4.5617
7	Root relative squared error	8.9515	54.8395	26.7982

Table III shows the Performance comparison of j48 algorithm with other tree based classification algorithm.

TABLE III

PERFORMANCE COMPARISON OF J48 ALGORITHM WITH OTHER TREE BASED CLASSIFICATION ALGORITHMS

Sr. No	**Parameter**	**J48**	**REPTree**	**Random tree**
1	Correctly classified instance	99.742	99.4959	9.665
2	Incorrectly classified instance	0.258	0.5041	0.335
3	Kappa statistics	0.9957	0.9916	0.9945
4	Mean absolute error	0.0003	0.0003	0.0003
5	Root mean squared error	0.0145	0.0198	0.0166
6	Relative_absolute_error	0.656	1.2601	0.5639
7	Root relative squared error	8.9515	12.2444	10.2236

Fig 2 represents weka screen shot for the performance of algorithm.

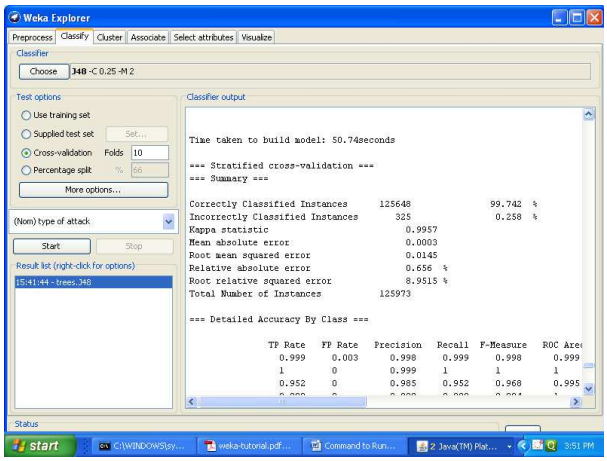

Fig 2

Algorithm performance

Analysis also suggests that true positive rate is high in tree based classification algorithm rather than rule based and bayes net. False positive rate of classification algorithm is lower than other algorithms. For performance comparison of the algorithm, attack class wise true positive and false positive rate are taken into account.

Based on experimentation it is suggested that decision tree are most suitable for intrusion attack classification. Decision trees represent a supervised approach to classification. Other reasons to choose decision tree methods are listed as follows:

- Decision trees are easy to understand.
- Decision trees can classify both categorical and numerical data, but the output attribute must be categorical.
- It makes no a priori assumptions about the nature of the data.
- Decision trees are easily converted to a set of production rules;

We briefly reviewed and implemented decision tree methods (i.e. REPTree, Random Tree,J48) and compared it with rule generation (OneR) and Bayesian(bayes net). Decision Tree algorithms shows better performance for intrusion attack classification. In our case, J48Tree shows the best performance.

Analysis of feature relevancy also performed in the dataset. 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 an 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. Table IV details the most relevant features for each class.

TABLE IV

MOST RELEVANT FEATURE PER CLASS LABEL

Feature No	Feature Name	Class
1	Duration	Normal
6	Dst_Bytes	
12	Logged_In	
15	Su_Attempted	
16	Num_Root	
17	Num_File_Creations	
18	Num_Shells	
19	Num_Access_Files	
31	Srv_Diff_Host_Rate	
32	Dst_Host_Coun	
37	Dst_Host_Srv_Diff_Host_Rate	
4	Flag	Smurf
25	Serror_Rate	
26	Srv_Error_Rate	
29	Same_Srv_Rate	
30	Diff_Srv_Rate	
33	Dst_Host_Srv_Count	
34	Dst_Host_Same_Srv_Rate	
35	Dst_Host_Diff_Srv_Rate	
38	Dst_Host_Serror_Rate	
39	Dst_Host_Srv_Serror_Rate	
2	Protocol_Type	Neptune
3	Service	
5	Src_Bytes	
23	Count	
24	Srv_Count	
27	Rerror_Rate	
28	Srv_Rerror_Rate	
36	Dst_Host_Same_Src_Port_Rate	
40	Dst_Host_Rerror_Rate	
41	Dst_Host_Srv_Rerror_Rate	
10	Hot	Back
13	Num_Compromised	
7	Land	Land
8	Wrong_Fragment	Teardrop
9	Urgent	Ftp_Write
11	Num_Failed_Logins	Guess_Pwd
14	Root_Shell	Buffer_Overflow
22	Is_Guest_Login	Warezclient

VIII. CONCLUSION

Data mining can improve intrusion based security attacks detection system by adding a new level of surveillance to detection of network data indifferences. It is highly required to identify appropriate features to categorize into different types of attack. Feature selection abbreviates the size of network data which improve finally performance of intrusion detection system. J48 algorithm which is used for experimentation is an efficient algorithm of classification in KDD cup dataset. Performance parameter demonstrated helps to improve efficiency of intrusion detection system.

REFERENCES:

[1] 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- 0535-8 .

[2] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second edition, 2005. Morgan Kaufmann.

[3] T. Lappas and K. P. , “Data Mining Techniques for (Network) Intrusion Detection System," January 2007.

[4] 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.

[5] Stephen Northcutt , Judy Novak “Network Intrusion Detection”, Third Edition, New Riders Publishing.

[6] 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.

[7] 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.

[8] 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/

[9] The KDD Archive. KDD99 cup dataset, 1999.

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

[10] Ghanshyam Prasad Dubey, Prof. Neetesh Gupta, Rakesh K Bhujade “ A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM”. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

[11] 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.

[12] 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)

[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] 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.

[15] Mark A. Hall, Geo_rey Holmes ”Benchmarking Attribute Selection Techniques for Discrete Class Data Mining” IEEE Transactions On Knowledge And Data Engineering, VOL. 15, NO. 3, MAY/JUNE 2003

[16] WEKA: Waikato environment for knowledge analysis . http://www.cs.waikato.ac. nz/ml/weka