Performance Analysis of Classifiers for Intrusive Data and Rough Sets Reducts

R.Ravinder Reddy CBIT
E. Padmalatha CBIT
Y.Ramadevi CBIT
K.V.N Sunitha BVRIT

Abstract
The fast change in the day to day activity need to analyze the intrusive data very accurately without losing performance. Intrusive behavior is critical for analyzing the data and Performance is crucial in the computational environment, when the user requires accuracy in the results. Finding intrusive behavior in the network with accuracy and speed is critical. In this paper we try to find out intrusive behavior more faster manner without losing the accuracy and improving the performance of the classifier. First we analyze the different characteristics of the classifiers and then we find out the reducts for the KDDCUP99 data set using the rough set theory and minimize the data set without losing any decisive attribute and prepare the new data set. With the new dataset we conducted the experiments for selected classification algorithms in data mining for both the datasets and compare performances.

Keywords
Data mining, classification, Rough sets, Intrusion detection

1. INTRODUCTION

As the network dramatically extended, security considered as major issues in networks. Cyber crime is increasing, and there have been various attack methods for the sensitive data, consequently. Intrusion systems have been used along with the data mining techniques like classification. Finding intrusion is important as well as with the accuracy and speed is required and marinating the critical characteristics of information like Integrity, Confidentiality and availability of the data.

Data mining researchers often use classifiers to identify important classes of objects within the data repository. An important component of many data mining projects is finding a good classification algorithm. Classification is particularly useful when a database contains examples that can be used as basis for future decision making; example for assessing credit card risk, intrusion detection etc. The classical scheme of knowledge discovery from data provided by Fayyad [2] in 1996 is simply called as data mining.

One of the most difficult tasks in the whole KDD process is to choose the right data mining technique [8] as the commercial software tools provide more and more possibility together and the decision requires more and more expertise on the methodological point of view. Here rough sets are introduced for enhancing the classifiers performance of the system.

2. CLASSIFICATION

Classification is one of the data mining and machine learning[6] technique used for classifies the objects into different classes based on the features of the class. Here the few classification algorithms are selected for the intrusive data for analysis of the intrusion detection.

2.1 Naive Bayes

The Naive Bayes [8] classifier provides a simple approach, with clear semantics, to representing and learning probabilitistic knowledge. It is termed naive because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process.

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Bayes' Theorem:
Prob(B given A) = Prob(A and B)/Prob(A)

2.2 J48 (C4.5 Decision Tree)

A decision tree is a 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.

Manuscript received August 5, 2014
Manuscript revised August 20, 2014

2.3 Lazy IBk

LBk [12] is a lazy classifier algorithm that makes use of the k-nearest-neighbor classifier. In this study, we choose the parameters for LBk as follow: k = 1; crossValidate = False; searchAlgorithm = LinearNNSearch; windowSize = 0.

2.4 Random forest

Random forests (RF) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

2.5 AdaBoost

AdaBoost is an algorithm for constructing a strong classifier as linear combination of simple weak classifiers. It has good generalization properties, feature selector with a principled strategy and sequential decision making

3. BASIC CONCEPTS OF ROUGH SETS

Rough set theory (RST)[4] is a useful mathematical tool to deal with imprecise and insufficient knowledge, find hidden patterns in data, and reduce dataset size (Pawlak, 1982; Komorowski, et al, 1998). Also, it is used for evaluation of significance of data and easy interpretation of results. RST contributes immensely to the concept of reducts. Reducts is the minimal subsets of attributes with the most predictive outcome. Rough Set is a machine learning method which generates rules based on examples contained within an information table. Rough set theory has become well established as a mechanism for solving the problem of how to understand and manipulate imprecise and insufficient knowledge in a wide variety of applications related to artificial intelligence.

Let K = (U,C) be an approximation space, where U is a non-empty, finite set called the universe; A subset of attributes $R \subseteq C$ defines an equivalent on U. Let [x]R ($x \in U$) denote the equivalence class containing x.

Given $R \subseteq C$ and $X \subseteq U$. X can be approximated using only the information contained within R by constructing the R-lower and R-upper approximations of set X defined as:

$RX = \{ x \in X \mid [x] R \subseteq X \}$

$RX = \{ x \in X \mid [x] R \cap X \neq 0 \}$ where

RX is the set of objects that belong to X with certainty, belong to X. The R-positive region of X is POSR(X) = RX

3.1 Dynamic Reducts

Knowledge reduct is an important step in knowledge discovery, and also a favourable method to extract the more generalized rules. Dynamic reducts can put up better performance in very large dataset, and also enhances effectively the ability to accommodate noise data. Dynamic reducts[5] are in some sense the most stable reducts of a given decision table, i.e. they are the most frequently appearing reducts in sub tables created by random samples of a given decision table. The set of decision rules can be computed from dynamic reducts in two different ways.

One can choose the best dynamic reducts and compute the decision rules using only attributes from the dynamic core i.e. from the union of these dynamic reducts.

Another possibility is to compute the set of rules separately for any of the chosen dynamic reduts and to create the union of the constructed decision rule sets

4. DATA SET

KDD'99 [3] has been the mostly wildly used data set for evaluation of anomaly detection methods. It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. In this paper we used the KDDCUP99 Dataset. In this we have taken 20% of the KDD test set for our experiment purpose it contains 11,580 instances. It contains 42 features for network intrusive data.

5. IMPLEMENTATION

The classifier predicts the class of each instance if it is a correct that is counted a success, if not it is an error. Here we have consider the different data mining classification algorithms for the experiment.

1) J48
2) Random Forest
3) Adaboost
4) Lazy IBK
5) Naive Bayesian Classifier

Before conducting the experiment preprocess the data for normalization and missing values. Then using the rough set method we calculated the dynamic reducts using genetic algorithm. The reduct set contains the 16 features, in the KDDCUP'99 dataset contains 42 features, the

rough set tool calculated the reducts for removing the redundancy and minimize the data set without loosing the decisive features. Using the rough set reduct calculated the new data set for faster analysis of the intrusive behavior.

The experiment is conducted for the above mentioned classification algorithms for the data set. And compared the results with the reduct dataset computations. There is huge performance achieved using the rough sets without losing the accuracy of the classifier.

These algorithms are run on the test mode k-fold cross-validation and observed the different performance measures listed below

1) Time taken to build the model
2) Correctly classified Instances
3) Incorrectly classified instances
4) Kappa statistics
5) ROC

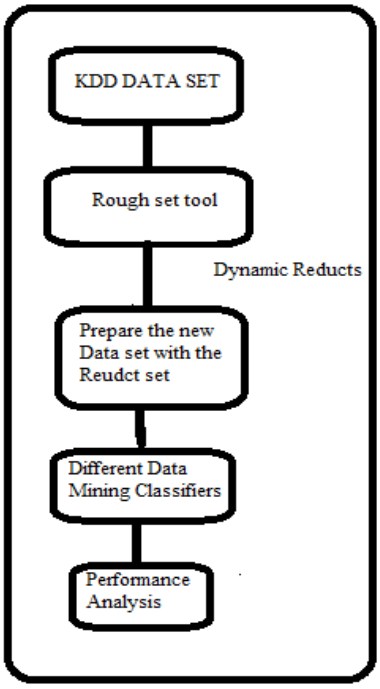


Fig 5.1 Process flow

6 RESULTS ANALYSIS

It's natural to measure a classifiers performance in terms of the error rate. Without an accuracy assessment, a classifier is just a pretty picture. In this paper we have consider the few classification algorithms for which we conducted the experiments, for every classification algorithm the performance measure are accuracy of the classifier and area under the curve(AUC) i.e, roc. We compared the different performance characteristics for both the techniques.

1. KDD Dataset
2. Reduct KDD Dataset

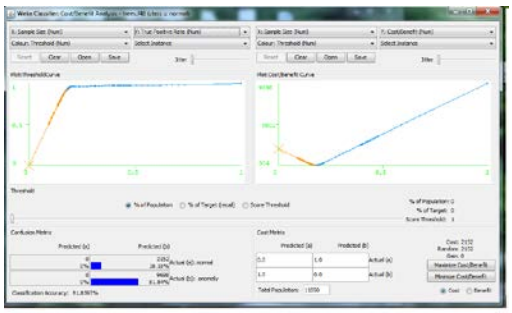
Different characteristics of the classifiers summarized in the fig 6.5 for both the methods and plot the graphs for the different parameters explained below

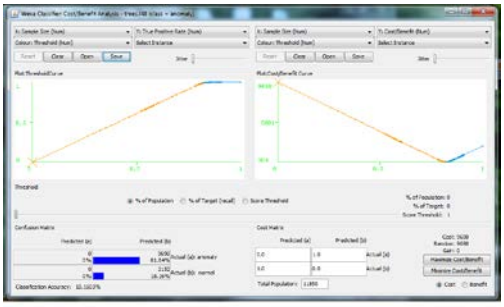
6.1 COST CURVES:

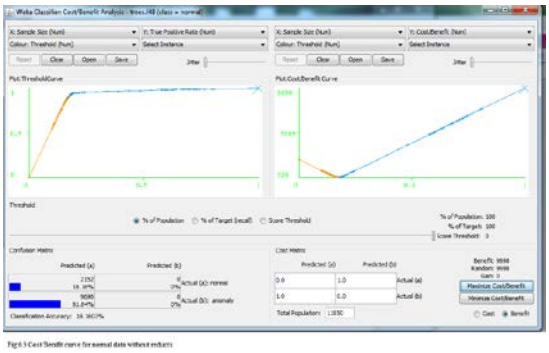
ROC curves and their relatives are very useful for exploring the tradeoffs among different classifiers over a range of costs. However they are not ideal for evaluating machine learning in situations with known error costs.

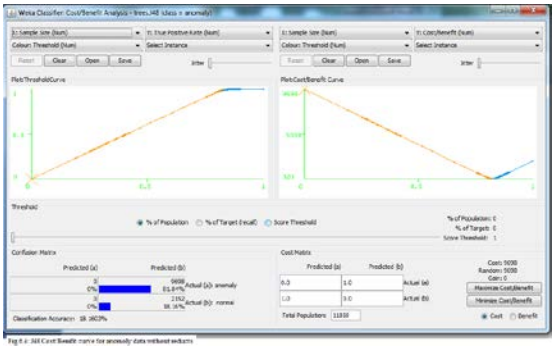
Cost curves [10] are different kind of display on which a single classification corresponds to a straight line that shows how the performance varies as the class distribution changes. One nice thing about cost curves is that the extreme cost values at the left and right sides of the graph are fp(false positives) and fn(false Negatives). Just as they are for the error curves. If the probability cost function exceeds about 0.45 and knowing the costs are could easily work out with this corresponds to in terms of class distribution. In situations that involve different class distributions, cost curves make it easy to tell when one classifier will outperform another. Cost curves can help to show which classifier to use when.

Here we have drawn the different cost curves for the J48 algorithm for KDDCUP99 dataset. In fig 6.1 drawn for the cost curve for normal data with the reducts. In fig 6.2 shows how the anomaly data cost curve. Likely in fig 6.3 and 6.4 depicts normal and anomaly cost curves for KDDCUP'99 Dataset without the reducts.









The crucial performance of the classification algorithm time required for building the model it shown in fig 6.6 for the rough set data and normal data, here the time taken for the classification of the rough set data is very faster than the normal dataset.

ROC curves and their relatives are very useful for exploring the tradeoffs among different classifiers over a range of costs. It shows in the fig 6.7 in this there is no much difference between these two modes. The classification accuracy is almost equal for the rough set data for the mentioned classifiers in the fig 6.5. Using the rough set data accuracy is very good.

Performance Measure / Classifiers	Time Taken to build the model		ROC		Correctly		In Correctly Classified		Kappa Statics	
	With Reducts Seconds	Without Reducts seconds	With Reducts	Without Reducts	With Reducts %	Without Reducts %	With Reducts %	Without Reducts %	With Reducts	Without Reducts
Ada Boost	15.2	34.57	0.907	0.925	90.3544	90.578	9.6456	9.646	0.6343	0.656
J48	4.51	12.16	0.9736	0.971	97.311	97.123	2.6667	2.667	0.9102	0.902
IBK	0.01	0.01	0.93	0.928	95.7131	95.747	4.2869	4.287	0.8549	0.856
Naïve Bias	1.54	3.46	0.82	0.845	77.7131	65.676	22.2814	22.284	0.3983	0.279
Random forest	8.92	10.06	0.993	0.993	97.4932	97.578	2.4932	2.493	0.9156	0.918

Fig 6.5 Comparison of classifiers with different parameters

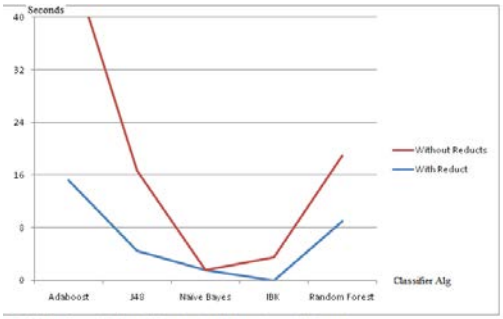


Fig 6.6: Time Comparisson for building the model for the reducts and without reducts

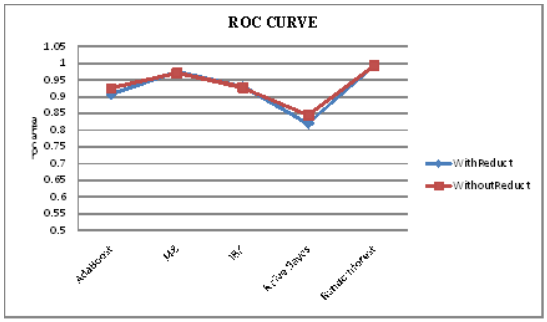


Fig 6.7 Roc curve for reduct data

The categories of the classification cannot reliably be distinguished by competent professionals. Kappa shows the classifiers assigning observations to classes.

Kappa was subsequently adopted by the remote sensing community as a useful measure of classification accuracy. In fig 6.8 the kappa statistics for the two different modes.

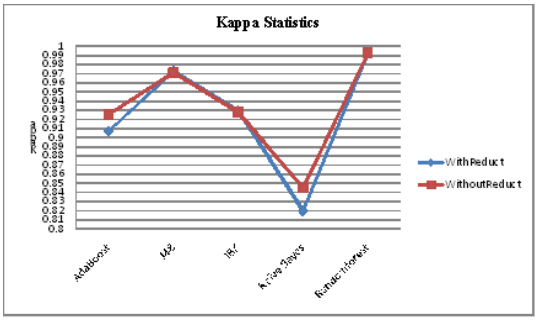


Fig 6.8 Kappa statistics for with and without reducts

7. CONCLUSION AND FUTURE WORK

The analysis on the different classifiers for intrusive data observed the rough set reducts have achieved the good performance on the data mining classifiers over the normal KDDCUP'99 data set. Accuracy of the classifiers are almost equal for both the reducts and without the reductts. Including the roc and kappa spastics are same in both cases

The methodology can be applied for all the classification algorithms for deciding which is the best classifier for intrusive data. Apply this model for real time data for generating the false alarams in effectively

REFERENCES

[1] vipin kumar10 classification algorithms performance analysis,springer,2006
[2] Usam M.Fayyad. Data mining knowledge discovery:Making sense out of data. IEEE Expert: Intelligent Systems and Their Applications,11(5):20-25,1996.
[3] KDD Cup 1999. Available on :http//kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,October 2007.
[4] Pawalak Z, "Rough sets[J]," International Journal of Computer and Information Sciences,vol.11,no. 5,PP.341-356,1982.
[5] Bazan J, Skowron A,Synak P, "Dyanamic Reducts as a Tool for Extracting Laws from Decision Tables in:Methodologies for Intelligent System,"Proc. 8Th International Symposium ISMIS'94, Charlotte, NG, October 1994, LNAI vol. 869, Springer Verlag 1994,346-355.
[6] Weka-Using data mining machine learning software,http://www.cs.waikato.ac.nz/ml/weka
[7] C. Jirapummin, N. Wattanapongsakorn, J. Kanthamanon, Hybrid neural networks for intrusion detection system, in: The International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Thailand, 2002, pp. 928-931.
[8] Karina Gibert,Miquel Sanchez-Marre,Victor Codina: Choosing the right data mining technique:Classification of methods and intelligent recommendation,Internation Environmental modeling and software society[IEMSS]
[9] Z. Pan, S. Chen, G. Hu, D. Zhang, Hybrid neural network and C4.5 for misuse detection, in: The 2nd International Conference on Machine Learning and Cybernetics, China, 2003, pp.2463-2467.
[10] book : Ian H, Witten & Eibe Frank, Data mining practical Machine Learning Tolls and Techniques.
[11] Abraham, R. Jain, Soft computing models for network intrusion detection systems, in: Knowledge Discovery, Computational Intelligence, vol. 4, Heidelberg, 2005, pp. 191-207.
[12] Aksoy,S.: K-Nearest Neighbor Classifier and Distance function s. Techical Report,Department of comuter EngineeringBilkent University(Feb 2008)