The International Journal Of Engineering And Science (IJES)
|| Volume || 4 || Issue || 7 || Pages || PP -54-60 || 2015 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

A Feature Selection Based On Rough Set For Improving Intrusion Detection System

[1]Sam Bordbar , [2]Mohd Tufik Bin Abdulah
[1]*Department of Computer Science, UPM University of Malaysia*
[2]*Department of Information Technology, UPM University of Malaysia*

ABSTRACT

Intrusion detection systems are monitoring a huge data set which consist of numerous records with various features and attributes in which some of them are redundant or do not have any effects on the intrusion detection systems. To cope with this serious problem, a feature selection approach based on the rough set theory is proposed to omit and eliminate the redundant and useless features that in this study are omitted by applying genetic algorithm, which is one of the main concepts of rough set theory. The performance of this approach is evaluated using standard NSL-KDD data set. The experimental results show that the rough set theory as a feature selection for intrusion detection system increased the accuracy and detection rate, and decreased the false alarm rate.

Date of Submission: 24 June 2015 Date of Accepted: 25 July 2015

I. INTRODUCTION

Intrusion detection system (IDS) can be categorized as a security element which is responsible for inspecting any inbound and outbound network intrusion activities against the PC or systems such as illegal access, misuses and any type of online hacker attacks.Firstly, it is the duty of IDS to gather all visitors or actions on the system. Afterwards, create the patterns and completely store it in the database. Next, all incoming visitors or actions with the stored patterns are checked and observed by IDS and then alarms are generated to declare risky actions to the administrator. Conventional security technology such as user authentication, data security, and firewalls which have been used as the first layer of computer protection, do not assure the security of system absolutely. For example, if the security password of one of the applications is poor and the user authentication method cannot prevent illegal access, then the system is highly at risk. Another example, misconfiguration and misinterpretation policies in Firewalls that cannot effectively protect the user from hackers access to the system. These traditional security systems are incapable of protecting systems against all missuses and harmful actions. Therefore, strike recognition program is required as an additional protection solution and supporting devices for defending systems against all types of attacks. Intrusion detection is useful not only for discovering successful attacks, but also for tracking and discovering any efforts to break the program protection.

Generally, intrusion detection systems can be classified into two groups: host-based IDS (HIDS) and network-based IDS (NIDS). As a matter of fact, assessing and managing all incoming and outgoing activities at one program section can be done and completed by Network-based intrusion detection system. To clarify it more, the assessment can be properly done by putting one catch tool (sensor) like one sniffer on the section checkpoint. This indicator can catch all program activities in this section and systematically and thoroughly evaluate whether these activities have been attacked or normal. The network-based IDS monitors packets to figure out whether they can be perfectly and precisely matched to predetermined styles or not. Moreover, based on the previous process, the attack identification program can find the attack and produce the alarm. Network-based IDS is capable of monitoring a large program environment with low cost for the program. In addition, Network-based IDSs can observe all slots for detrimental activities such as http port 80.

On the other hand, host-based intrusion detection system (HIDS) examines all system traffic and monitors what happens to each variety independently. More detailed, the HIDS is capable of discovering risky actions and normal functions by obtaining logging information or tracking the variety utilization faster (De Lima, Degaspari and Sobral, 2008). Host-based IDS is located on a single variety or in a computer system and observe and examine only one program. More specifically, the work of this program work depends on customers' behaviors.In this case, all the behaviors of the customers can be gathered and stored as a regular behavior. To discover the strike, the intrusion detection system observes all actions in the program and determines any differences discrepancies from the regular behavior.

There is another classification of intrusion detection system based on detection method: misuse (signature based) detection and anomaly detection. Signature based IDS contains a database of known weaknesses. It also monitors traffic and looks for a signature or a pattern coordinate. In other words, it functions as the same way as a malware scanning device, by looking for a known identification or trademark for each particular attack occasion. It can be placed on a system to look at the system weaknesses or can be placed on various parts of the system. The signatures and styles are used to recognize strikes involved in the various areas of a system such as resource and location.

Anomaly based IDS is also known as Heuristic or Behavior based. In this context, anomaly based IDS examines the traffic styles and figure out of regular actions. After that, it applies mathematical or heuristic actions of the event to discover if the data coordinate with a regular behavior or not. Activities here which do not coordinate with the approved regular patterns are known as strikes.

False Positive and False Negative are two typical significant terms in all intrusion detection systems. False positive is a condition where IDS increases alert of strikes for something that is not really an attack. It is of vital importance to note that if too many false positives happen, it would make the administrator be less assured about the alarm system and hence, there would be a probability of neglecting the actual strike by the administrator. This is a typical issue which happened in all IDS and most of the researchers in IDS area are trying to decrease false positives and improve detection rate which is one of the primary objectives of IDS.A false negative on the other hand, is a condition where IDS does not improve an alert for an actual strike. This implies an actual strike that can happen without the counter actions being taken. This must be highly risky for the program, because the actual strikes are completely neglected.

The point here is that the dimensionality of the problem may be considerably higher. At the time of data collection, this process can be aggravated which is because of the fact that it is often mistakenly accepted that having more features can be equal with being more knowledgeable.Thus, increasing the likelihood of having enough information to distinguish between both classes of these data can exist.IDS system works on big datasets with a large number of objects and different features. Furthermore, computational explosion can result from the High-dimensional data sets that in fact it can create various difficulties regarding search space, and also it can be shown as an extra dimension (feature) in which all of them can make this problem, computational explosion, more complex and serious. Working on these datasets needed a considerable amount of time and huge amount of memory. Some of these attributes however, do not have any major effects on security issues and IDS results. It has been always highly complicated to find an algorithm to do the right feature selecting that can effectively eliminate these redundant features.

In this work we proposed a feature selection system that is capable of greatly reducing the false alarm rates and markedly increasing the accuracy of the detection system. More importantly, the main goal of this research is to propose a feature selection system by using rough set theory which has far better results than using the IDS without the rough set theory. The genetic algorithms were used to design and apply the reduct.

A. Rough Set Theory

Pawlak (1982) proposed a specific theory named as rough set theory (RST). The main idea behind this method is the calculation of upper and lower boundaries of a set. Rough sets are the sets defined through these upper and lower approximation. Rough Set Theory has turned into an important and sophisticated device in the determination of a wide range of various issues such as showing uncertainty or general knowledge, knowledge discovery, estimation of the quality and availability of information with respect to the existence of a note of date patterns, evaluation and identification of date dependency. The degree of rough set systems which have been used today is much more extensive than before.

B. Attribute reduction

It is frequently required to maintain a concise form of the information in the system, but there are data which can be omitted, without changing the main properties and the uniformity of the system. Therefore, the new dataset has the same approximation precision and the same dependency grade with the original set of attributes.However, with one difference, the set of attributes which needed to be considered will be fewer. Decreasing a data framework such that the set of attributes of the reduced information system is autonomous and no attribute can be omitted without losing some information from the system, the result of this procedure is known as reduct (Rissino and Lambert-Torres, 2009). Reducts do not have any unnecessary attributes. Consequently, the reducts have the capability to efficiently organize data, without loosing any knowledge.

II. LITERATURE REVIEW

The idea behind the search for valuable patterns in data has had numerous names in the past like knowledge extraction, data mining, and data pattern discovery. At a fundamental level knowledge discovery from data is the development of approaches and procedures which ‘make sense of data’. The main issue with the knowledge discovery process is attempting to address the mapping low-level data into other forms which may be more compressed, more humanly interpretable, or more beneficial.

The traditional manual procedure of changing data into knowledge (Fayyad, Piatetsky-Shapiro and Smyth, 1996) depends on human analysis and interpretation. For example, in market research, it is very common for commercial organizations to periodically employ an expert to analyze data relating to current consumer trends and poor chasing habits. Then exerts compile an assessment, detailing the analysis, and present it to the organization. This assessment forms the basis for future marketing strategies, and decision-making with regard to the target audiences. In criminal forensic case, investigators need to investigate huge amounts of scientific evidence, carefully analyses and cataloging objects, material fragments, and fingerprints. Before generating plausible scenarios or events.

Whether in science, crime detection, finance, machine performance, or any other area, this traditional method for data analysis highly depends on the fact that at least one expert must have a detailed understanding of the data. As a result, the method of manual manipulation of data is very time-consuming, costly, and of course very subjective. As the volume of data grows intensely, manual data analysis is becoming totally unreasonable in some fields. Data is growing in terms of the number of data objects and the number of features. Clearly, these types of knotty problems are beyond the scope of the human being, and hence, such analysis requires automation. Today's data not only have a large number of objects, but also there may be a great number of features. It is a problem that often frustrates the applications of machine learning techniques for knowledge discovery. Solutions to this big problem contain approaches which decrease the overall dimensionality of the data. Such approaches are known as dimensionality reduction techniques (Lee and Verleysen, 2007).

Gong, Gong and Bi (2011) have proposed a feature selection based on Genetic Quantum Particle Swarm Optimization (GQPSO). In this paper, they proposed GQPSO algorithm which is based on QPSO algorithm and Genetic algorithm. This technique decreases out of work and unrelated features. Nguyen and De la Torre (2010) used a k-nearest neighbor (KNN) classifier as fitness purpose of the Genetic algorithm and also as a classifier. Being easy to calculate the weight of the features is the KNN advantage. Experimental results indicate an increase in intrusion detection accuracy. Xing, Jordan and Karp (2001) have suggested a method based on genetic search methods. A novel artificial Intelligence paradigm called the artificial immune system (AIS) was proposed. A genetic search approach was used for correlation based feature selection. The experimental outcomes confirm recall of 99.7% of normal data. Sridevi and Chattemvelli (2012) have proposed a novel feature selection method for determining an optimal feature set. The filter is a Correlation-based Feature Selection (CFS). It evaluates the advantages and merits of the feature subset. This hybrid feature selection technique decreased the computational source while keeping the detection and false positive rate unacceptable boundaries.

III. METHODOLOGY

A reduct is a set of attributes that saves the essential and vital characteristics of the original data set. Therefore, the attributes that do not belong to a reduct are worthless (Rissino and Lambert-Torres, 2009) This study proposes a feature selection based on reduct by using Genetic algorithm. The methodology used in this study divided into three phases.

Phase1: Dataset

NSL-KDD dataset was downloaded and used from the http://nsl.cs.unb.ca/NSL-KDD/ . The available records in the dataset are raw data. Thus, we divided each record and its attributes and formed a table where rows represent attributes and columns represent records.

Phase 2: Design

In this phase, the reduct is applied to the new tables from phase1 by using the genetic algorithm based on the rough set model. The result of the reduct is a subset which its attributes are the features for feature selection.

Phase 3: Test Evaluation

In this phase, the effect of our feature selection on the intrusion detection was checked and compared with other feature selection algorithms.

A. Dataset Description

One of the greatest difficulties in network-based intrusion detection is the comprehensive amount of information gathered from the system. Therefore, before providing the information to a machine learning criteria, raw system traffic should be described into higher-level activities such as relationship information.NSL-KDD intrusion detection dataset, which is based on DARPA 98 dataset, provides labelled information for scientists who work in the area of intrusion detection, and is the only labelled dataset openly available.This dataset is divided into two sets: testing set and training set.Each packet in NSL-KDD dataset consists of 41 features that derived from each connection.

B. RSES

There are some simulators in the case of rough set theory studies such as Rosetta and RSES. In this study, RSES is used to calculate genetic reduct based on the rough set model.

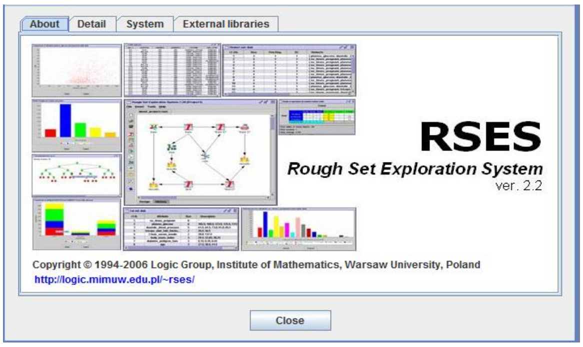
About
Detail
System
External libraries
RSES
Rough Set Exploration System
ver. 2.2
Copyright © 1994-2006 Logic Group, Institute of Mathematics, Warsaw University, Poland
http://logic.mimuw.edu.pl/~rses/
Close

Fig. 1. RSES simulator

C. Weka

There are several simulators in the case of machine learning studies such as Matlab and Weka. In this study, Weka simulator is used to implement intrusion detection systems.

IV. EXPERIMENTAL RESULTS AND EVALUATION

The NSL-KDD divided to KDD train and KDD test. After calculating the reduct using geneticalgorithm for these datasets, the redundant attributes omitted and the remained features were used as feature selection. Figure 1 shows the KDD test reduct result and Figure 2 KDDtrain reduct result.

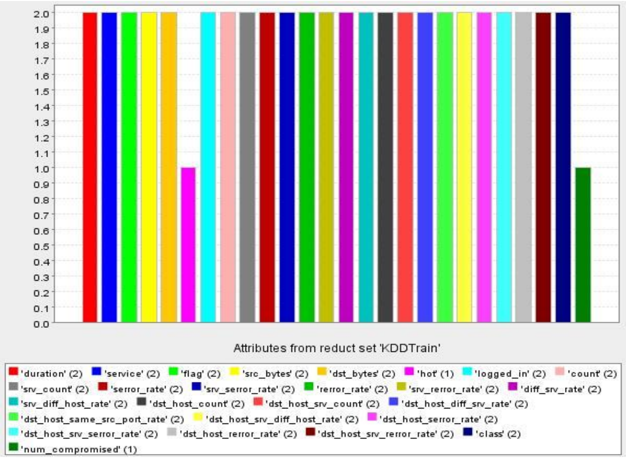

Fig. 2. KDDtrain reduct result

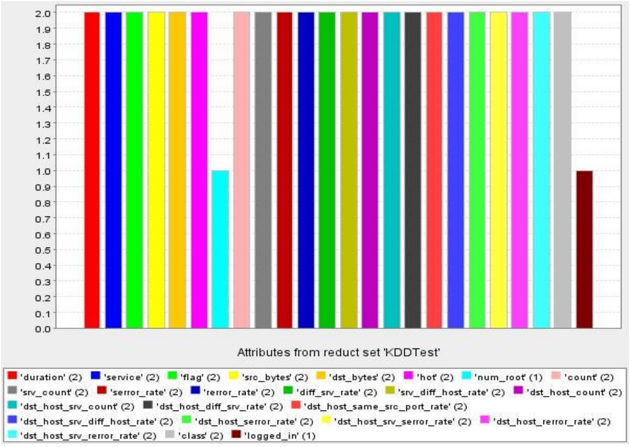

Fig. 3. KDDtest reduct result

These attributes were used as features for the IDS and Naïve-Bayes was used as the classifier in our IDS.

Three different experiments have been carried out in this research to compare the major effects of different feature selection algorithm on IDS systems. First, The Naive-Bayes classification is selected as a single classifier without using any feature selection. Secondly, Dimension Reduction in Intrusion Detection Features

Using Discriminative Machine Learning Approach (Bajaj and Arora, 2013) is used as a feature selection algorithm. Third, Rough Set Theory is used as feature selection. Evaluation of Rough set reduct with Naive-Bayes approach is based on the comparison with the following approaches:
1. Single Naive-Bayes classifier without feature selection
2. Discriminative Machine Learning as a Feature selection + Naive-Bayes as a classifier
In terms of accuracy, detection, and false alarm rate based on the flow chart in figure 4.

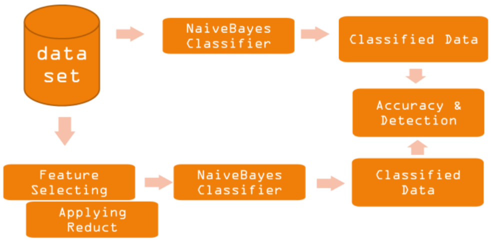

Fig. 4. Methodology flow chart

Table 1 shows the output of the simulated intrusion detection system by using KDDtrain dataset and Table 2 shows the output of simulated intrusion detection systems by using KDDtrain.

TABLE I
RESULTS FOR KDDTRAIN DATASET

KDDtrain	Accuracy	Detection rate	False alarm rate
NaiveBayes	90.3829%	92.22%	6.35%
FS(Discriminative)+NaiveBayes	90.444%	92.39%	6.20%
RST + NaiveBayes	90.4528%	93.17%	5.44%

TABLE II
RESULTS FOR KDDTEST DATASET

KDDtest	Accuracy	Detection rate	False alarm rate
NaiveBayes	80.73%	94.86%	5.00%
FS(Discriminative)+NaiveBayes	80.73%	94.87%	4.99%
RST + NaiveBayes	81.26%	95.00%	4.92%

The proposed feature selection approach and the other two approaches are compared in terms of accuracy, detection and false alarm rate. Figure 5 shows the accuracy of the three simulations for both KDDtrain and KDDtest datasets.

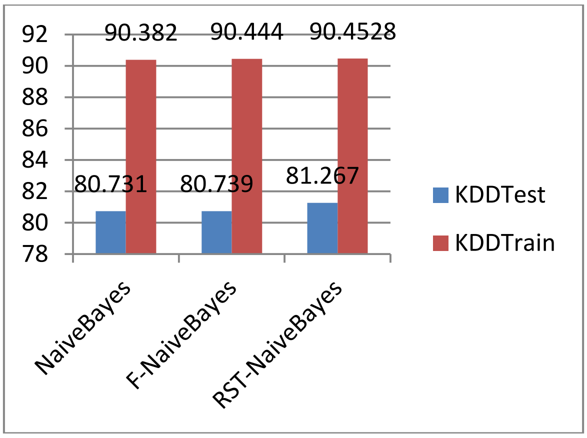

Fig. 5. Accuracy rate

The accuracy results for KDDtest dataset show a 0.53 % improvement compared to the discriminative feature selection and 0.53% compare to solo IDS without feature selection.

The detection rate is defined as the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances present in the test set. Figure 6 shows the detection rate of the three simulations for both KDDtrain and KDDtest datasets. Results show improvement compared to other approaches.

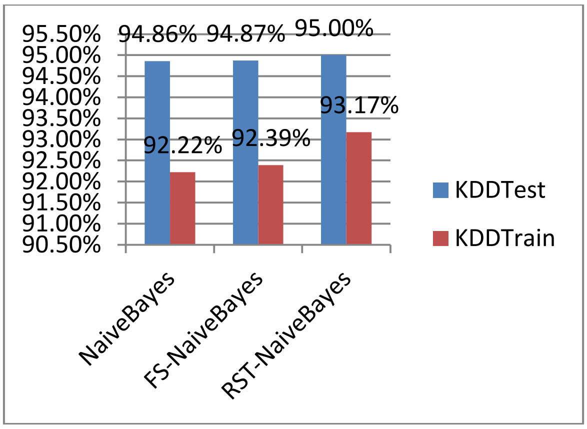

Fig. 6. Detection rate

The detection results for KDDtest dataset show a 0.13 % improvement compared to the discriminative feature selection and 0.14% compare to solo IDS without feature selection.

False alarm rate defined as the number of 'normal' patterns classified as attacks (False Positive) divided by the total number of 'normal' patterns. Figure 7 shows the false alarm rate of the three simulations for both KDDtrain and KDDtest datasets. Results show a decrease of false alarm rate compare to other approaches.

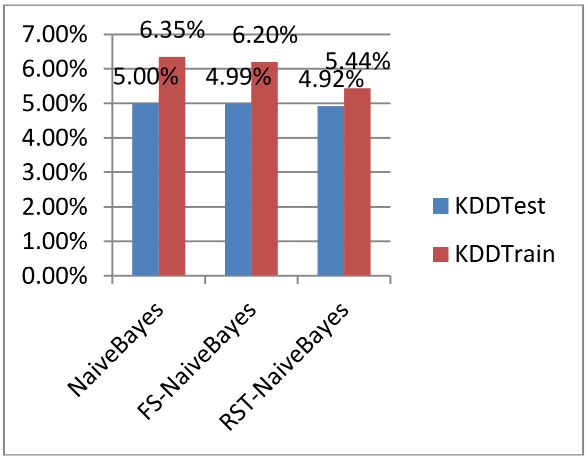

Fig. 7. False alarm rate

The results for KDDtest dataset show a 0.07 % improvement compared to the discriminative feature selection and 0.08% compared to solo IDS without feature selection.

V. CONCLUSION AND FUTURE WORK

Due to the related work to Feature selection and IDS, the challenges and gaps in this area are finding a feature selection approach to omit redundant attributes, save memory, increase accuracy and detection rate while decreasing the false alarm rate. In this work RST reduct proposed as a feature selection approach. A Genetic algorithm was used to apply the reduct. For performing a comparison, three IDS’s are designed, solo IDS without feature selection, IDS hybrid with a discriminative machine learning approach as feature selection, and IDS hybrid with a RST approach as feature selection. In comparison, of the three approaches, proposed approach, showing a 0.78 % improvement in detection rate compared to discriminative machine learning approach and also 0.98% improvement compared to solo IDS system ,while the false alarm rate decreases 0.76% compared to the discriminative machine approach and 0.91% compared to solo IDS system. In addition, proposed approach usese less memory compared to solo IDS system and discriminative machine learning approach due to omitting redundant data in reduct.

Future works will be focused on finding the significant and profound effects of rough set theory as a feature selection for Cloud systems. This work requires a new data set recorded from the Cloud networks. We will try to propose an intrusion detection for Cloud systems, using rough set theory as feature selection.

REFRENCES

[1] Bajaj, K. and Arora, A. (2013)' Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach.' *International Journal of Computer Science Issues (IJCSI), 10* (4).

[2] De Lima, I. V. M., Degaspari, J. A. and Sobral, J. B. M. (2008) 'Intrusion detection through artificial neural networks.' *In Network Operations and Management Symposium*, 867-870.

[3] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases.' *AI magazine*, 17(3), 37.

[4] Gong, S., Gong, X. and Bi, X. (2011) 'Feature selection method for network intrusion based on GQPSO attribute reduction.' In *Multimedia Technology (ICMT), 2011 International Conference on* (6365-6368).

[5] Lee, J. A. and Verleysen, M. (2007) *Nonlinear dimensionality reduction*. Springer.

[6] Nguyen, M. H. and De la Torre, F. (2010) 'Optimal feature selection for support vector machines.' *Pattern recognition*, 43(3), 584-591.

[7] Pawlak, Z. (1982) Rough sets. *International Journal of Computer & Information Sciences*, 11(5), 341-356.

[8] Rissino, S. and Lambert-Torres, G. (2009) *Rough Set Theory–Fundamental Concepts, Principals, Data Extraction, and Applications. Data Mining and Knowledge Discovery in Real Life Applications, J. Ponce and A. Karahoca (Eds.),* InTech Publishers.

[9] Sridevi, R. an Chattemvelli, R. (2012) 'Genetic algorithm and artificial immune systems: A combinational approach for network intrusion detection.' *In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (494-498).

[10] Xing, E. P., Jordan, M. I. and Karp, R. M. (2001) 'Feature selection for high-dimensional genomic microarray data.' In *ICML*, 1, 601-608.