Rough Set Significant Reduct and Rules of Intrusion Detection System

Noor Suhana Sulaiman[1], Rohani Abu Bakar[2],
[1] Faculty of Computer, Media and Technology Management
TATi University College
24000 Kemaman, Terengganu, MALAYSIA
e-mail: suhana@tatiuc.edu.my
[2] Faculty of Computer Systems & Software Engineering
Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300 Kuantan, Pahang, MALAYSIA
e-mail:rohani@ump.edu.my

ABSTRACT: *Intrusion Detection System deals with huge amount of data which contains irrelevant and redundant features causing slow training and testing process, also higher resource consumption as well as poor detection rate. It is not simply removing these irrelevant or redundant features due to deteriorate the performance of classifiers. Furthermore, by choosing the effective and important features, the classification mode and the classification performance will be improved. Rough Set is the most widely used as a baseline technique of single classifier approach on intrusion detection system. Typically, Rough Set is an efficient instrument in dealing with huge dataset in concert with missing values and granularing the features. However, large numbers of generated features reducts and rules must be chosen cautiously to reduce the processing power in dealing with massive parameters for classification. Hence, the primary objective of this study is to probe the significant reducts and rules prior to classification process of Intrusion Detection System. All embracing analyses are presented to eradicate the insignificant attributes, reduct and rules for better classification taxonomy. Reducts with core attributes and minimal cardinality are preferred to construct new decision table, and subsequently generate high classification rates. In addition, rules with highest support, fewer length, high Rule Importance Measure (RIM) and high coverage rule are favored since they reveal high quality performance.* The results are compared in terms of the classification accuracy between the original decision table and a new decision table.

Keywords: Rough Set, Reducts, Rules, Classification, Intrusion Detection System

Received: 19 October 2016, Revised 2 December 2016, Accepted 11 December 2016

© 2017 DLINE. All Rights Reserved

1. Introduction

Electronic commerce and the recent online consumer boom have forced a change in the basic computer security design for systems on a shared network. This is because of the systems are now designed with more flexibility and less barrier security. The combination of user friendliness and public accessibility, although advantageous for the average person, inevitably renders any exchanged information vulnerable to criminals. Consumer information, employee data or intellectual property stored in

internal data warehouses are all at risk, from external attackers and disgruntled employees who might abuse their access privileges for personal gain. Security policies or firewalls have difficulty in preventing such attacks because of the hidden weaknesses and bugs contained in software applications [1]. Moreover, hackers constantly invent new attacks and disseminate them over the Internet. Disgruntled employees, bribery and coercion also make networks vulnerable to attacks from the inside.

An intrusion is defined to be a violation of the security policy of the system. Intrusion detection thus refers to the mechanisms that are developed to detect violations of system security policy. Thus, Intrusion Detection System is significant in network security. IDS becomes a hot topic in recent years, it detects and identifies intrusion behavior or intrusion attempts in a computer system by monitoring and analyzing network packet in real time. It is an effective tool for determining whether unauthorized users are attempting to access, have already accessed or have compromised the network. Consequently, it is important to find out intrusion quickly and effectively. IDSs may be some software or hardware systems that monitor the different events occurring in the actual network and analyze them for signs of security threats.

IDS consist of two approaches which are anomaly and misuse detection. The idea of misuse detection is to represent attacks in the form of a pattern or a signature so that the same attack can be detected and prevented in the future. These systems can detect many or all known attack patterns, but they are of little use for detecting naïve attack methods [2]. Pattern-matching solutions primarily use misuse detection. They employ a library of signatures of misuse, which are used to match against network traffic. The idea of anomaly detection is to build a normal activity profile for a system. The main advantage with anomaly intrusion algorithms is that they can detect new forms of attacks, because these new intrusions will probably deviate from the normal behavior [3, 4]. Anomalous activities that are not intrusive are flagged as intrusive, though they are false positives. Actual intrusive activities that go undetected are called false negatives.

Rough set is a fairly useful intelligent technique that has been applied to the IDS domain and is used for the discovery of data dependencies, evaluates the importance of attributes, discovers the patterns of data, reduces all redundant objects and attributes, and seeks the minimum subset of attributes. In rough information system, there often exist some condition attributes that do not provide any additional information about the objects. Consequently, those attributes should be reduced if those condition attributes are eliminated. A decision table may have more than one reduct. Any decision table can be used to replace the original table. Furthermore, the best reduct with the most minimal number of attributes criteria should be consider if there are two or more reducts with the same number of attributes occurred. If this case happened, then the reducts with the least number of combinations of values attributes is selected. The rules derivations are generated from the reduct. There are possible to have a lot of rules to use in classification process.

Many analyses have been done to come out with significant rules. Since the generated rules possible to have in large number of rules, it is important to know whether all rules played a role in the classification process. Indranil Bose, [5] has suggested that, to find the most significant rules for each sample, the rules are sorted according to the value of their support. The generated rules do not differ much in terms of length and thus support is used as the criterion for ranking the rules. Subsequent analysis is the evaluation of the rules length to obtain testing accuracy. Typically, rules of less length ascend to a higher overall testing accuracy. This indicates that the dataset led to the formation of a smaller number of rules, can correctly recognize the problem. The overall testing accuracy is highest when the training sample is reduced to 10% from the original sample. Then, the experiment of changing the parameters associated with the testing procedure is implemented. The experiment is conducted using four factors; balance of sample, ratio of training to testing sample size, testing sample size and training sample. The experiment resultant that there was no significant effect of changing the balance of the sample, training to testing sample size, training sample size and testing sample size on testing accuracy across all samples. The best classification result is obtained and the comparisons are made with two other statistical approaches; logistic regression and discriminant analysis. The reported results reveal that RST method generally performed better than the others in terms of classification accuracy on training and testing samples.

Not all the rules are significant for better classification. Hence, significant reducts and rules must be chosen to contribute better classification. Subsequently, core attributes and minimal cardinality are exploited to expedite the significant reducts and the rules measurement of rules support, rules length, rules coverage and rule important measure (RIM). The results show that the proposed method give better classification performance compared to standard rough set in terms of accuracy.

2. Rough Set Theory

Rough-Set Theory (RST) was introduced by Polish logician, Professor Zdzis³aw Pawlak in 1982 to cope with imprecise or vagueconcepts. Recently, it is one of the most developing soft computing methods for the identification and recognition of common patterns in data, especially in the case of uncertain and incomplete data. The mathematical foundations of this method are based on the set approximation of the classification space [6,7]. A rough set is a formal approximation of a crisp set which is *conventional set*, in terms of a pair of sets which give the *lower and the upper approximation* of the original set [8].

Knowledge base for rough set processing is stored as a table containing conditional and decision attributes. A method of knowledge representation is very important for Rough-Set data processing. Data are stored in a decision table. The columns represent attributes and the rows represent objects whereas every cell contains attribute value for corresponding objects and attributes. Decision tables are also called information systems. A decision table (*DT*) is the quadruple $T = (U, A, C, D)$, where U is a nonempty finite set of objects called the universe, A is a nonempty finite set of primitive attributes, and $C, D \subseteq A$ are two subsets of attributes that are called the condition and decision attributes.

A unique feature of the RST method is its generation of rules that played an important role in predicting the output. Rosetta listed the rules and provides some statistics for the rules which are support, accuracy, coverage, stability and length. Below is the definition of the rule statistics [9].

i) The rule LHS support is defined as the number of records in the training data that fully exhibit property described by the IF condition.

ii) The rule RHS support is defined as the number of records in the training data that fully exhibit the property described by the THEN condition.

iii) The rule RHS accuracy is defined as the number of RHS support divided by the number of LHS support.

iv) The rule LHS coverage is the fraction of the records that satisfied the IF conditions of the rule. It is obtained by dividing the support of the rule by the total number of records in the training sample.

v) The rule RHS coverage is the fraction of the training records that satisfied the THEN conditions. It is obtained by dividing the support of the rule by the number of records in the training that satisfied the THEN condition.

vi) The rule length is defined as the number of conditional elements in the IF part.

3. Rough Set Methodology for Classifying IDS Dataset

When a classifier is presented with a new case, the rule set is scrutinized to find pertinent rule that is the rules that the predecessors match the case. If no rule is found, the most frequent outcome in the training data is chosen. If more than one rules match, these may in turn indicate more than one possible outcome. A voting process is executed across the matched rules to resolve the conflicts and to rank the predicted outcomes.

This study employed KDD Cup (1999) dataset which is an IDS benchmark dataset. The dataset was prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Lab. The database contained a wide variety of intrusions simulated in a military network environment. The data are labeled as attack or normal, and furthermore are labeled with an attack type that can be grouped into four broad categories of attacks. Some intrusion experts suggested that most novel attacks are variants of known attacks, and the signature of known attacks can be sufficient to catch novel variants. The original data contains 744MB data with 4,940,000 records. Based on Table I, each TCP connection has 41 features with a label which specifies the status of a connection as either being normal, or a specific attack type. Attacks in the data sets are divided into four main categories: [10]

*** DOS (Denial of Service):** Such as ping of death attack and syn flood

*** U2R (User to Root):** Such as eject attack; unauthorized access to root privileges). unauthorized access to local superuser (root) privileges, e.g., various buffer overflow attacks

*** R2L (Remote to local):** Such as guest attack; unauthorized access from a remote machine unauthorized access from a remote machine, e.g., guess_passwd

*** PROBING:** Such as port scanning attack. surveillance and other probing) surveillance and other probing, e.g., port scanning.

The major objectives performed by detecting network intrusion are stated as recognizing rare attack types such as U2R and R2L, increasing the accuracy detection rate for suspicious activity, and improving the efficiency of real-time intrusion detection models. This detects that the training dataset consisted of 494,019 records, among which 97,277 (19.69%) were 'normal', 391,458(79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R attacks.

No	Features	No	Features
1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	same_srv_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login	42	normal or attack

Table 1. KDD Cup 99 Features

4. Generation of Significant Reduct and Rules

The significant reduct are based on core attributes and minimal cardinality. The core attributes is the set of attributes which is common to all reducts. The core is the set of attributes which is possessed by every legitimate reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence class structure. The intersection of all reducts is called the core reduct; the elements of attributes that cannot be eliminated as well as the set all indispensable attributes. The core is defined as;

$$\text{Core}(C) = \cap \text{Red} \tag{1}$$

In rough set theory, all of indispensable attributes should be restricted in an optimal attribute subset.Core is the set all indispensable attributes. The process of searching indispensable attributes is that of finding the CORE.

In rough set attribute reduction, a reduct with minimal cardinality is searched for. An effort is made to locate a single element of the minimal reduct set $\text{Red}_{\min} \subseteq \text{Red}$ The reduct with minimal cardinality is the reduct with minimal length.

$$\text{Red}_{\min} = \{R \in \text{Red} \mid \vee A^{'} \in \text{Red}, |R| \leq |R^{'}|\} \tag{2}$$

The significant rules are based on rules support, rules length, rule coverage, rule accuracy and rule important measure (RIM). Given a description contains a conditional part α and the decision part β, denoting a decision rule $\alpha \rightarrow \beta$. The support of the pattern α is a number of objects in the information system A has the property described by α.

$$\text{Sup } port(\alpha) = ||\alpha|| \tag{3}$$

The support of β is the number of object in the information system A that have the decision described by β.

$$\text{Sup } port(\beta) = ||\beta|| \tag{4}$$

The support for the decision rule $a \rightarrow b$ is the probability of that an object covered by the description is belongs to the class.

$$\text{Sup } port(\alpha \rightarrow \beta) = \text{Sup } port(\alpha \,.\, \beta) \tag{5}$$

For the accuracy measurement, the quantity accuracy $(\alpha \rightarrow \beta)$ gives a measure of how trustworthy the rule is in the condition β. It is the probability that an arbitrary object covered by the description belongs to the class. It is identical to the value of rough membership function applied to an object x that match α. Thus accuracy measures the degree of membership of x in X using attribute B.

$$\text{Accuracy } (\alpha \rightarrow \beta) = \frac{\text{Sup } port(\alpha \,.\, \beta)}{\text{Sup } port(\alpha)} \tag{6}$$

Coverage measurement measures the behavior of pattern a in describing the decision class defined through b. It is a probability that an arbitrary object, belonging to the class C, and is covered by the description D.

$$\text{Coverage } (\alpha \rightarrow \beta) = \frac{\text{Sup } port(\beta)}{\text{Sup } port(\alpha \,.\, \beta)} \tag{7}$$

The rules are said to be completed if any object belonging to the class is covered by the description coverage is 1, while deterministic rules are rules with the accuracy is 1. The correct rules are rules with both coverage and accuracy is 1.

Rules generated from reduct are representative rules extracted from the data set. Since a reduct is not unique, rule sets generated from different reducts contain different sets of rules. However, more important rules will appear in most of the rule sets. Less important rules will appear less frequently than those more important ones. Some rules are generated more frequently than the others among the total rule sets. Such rules are considered as more important rules. The Rule Importance Measure (RIM) [11] is computed according to the frequency of an association rule among the rule sets. The Rule Importance Measure (RIM) is defined as follows,

$$\text{Rule Importance Measure} = \frac{\text{Frequency of Appeared Rules from Reduct Set}}{\text{Number of Reducts Set}} \tag{8}$$

In this study, the following tasks have been done by using KDD Cup 99 dataset:

i) Missing values of the dataset have been removed by incorporating the incomplete process.

ii) Dataset are split into 70% of training and 30% of testing records

iii) Training dataset has gone through the Equal Frequency Binning Discretization.

iv) The discretization data generates the reducts.

v) The rules are produced for classification process.

vi) An analysis and evaluation of:

a) The generated reduct by

i. Choosing minimal cardinality.

ii. Core attributes in the generated reduct are analyzed.

b) A new decision table is constructed based on the attributes consist of reducts with minimal cardinality and core attributes

vii) New decision table have the same process as no i) to v)

viii) Second phase of analysis and evaluation:

a) Generated Rules are analyzed accordingly;

i) Rules with highest support values are chosen

ii) Rules with less length are preferred

iii) Rules with coverage of 1

iv) Rules with accuracy of 1

v) Rules with highest percentage of RuleImportance Measure (RIM) are favored

ix) Testing dataset has gone through the Equal Frequency Binning Discretization.

x) Classification process is done as in step (vi)

xi) Results of classification are compared in terms of classification accuracy of original decision table and new decision table.

xii) Experimental results are analyzed and discussed.

5. Experimental Result and Discussion

The dataset in this study contain randomly generated 29,995 records having 41 features. The data are divided into two parts; training and testing group. The training group is split into 70% which equal to 20,996 records, while the testing group is accounted for 30% which equal to 8,999 records. The classification is implemented using standard voting classifier. The derived rules from the training phase are used to test the effectiveness of the unseen data.

Training data is discretized using EFB to obtain an equal number of objects into each interval. Genetic Algorithm is used for reduct generation as it provides more exhaustive search of the search space. Reduct with object related is used, which produce a set of decision rules or general pattern through minimal attributes subset that discern on a per object basis. The reduct with object related have capability in generating reduct based on discernibility function of each object.

Based on the generated reducts with length 13,14 and 15, the core attributes are {*service, src_bytes, dst_bytes, hot, count, srv_count, srv_diff_host_rate, dst_host_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate,dst_host_srv_diff_host_rate*}. These attributes are important attributes to obtain better classification in testing phase. The reduct with minimal cardinality also contribute to the connotation reduct in generating the significant rule. It will consider the reduct with minimal cardinality of minimal length. In this experiment, the reduct with minimal cardinality is {infectious endocarditis} with length of 13. Based on these core attributes and attributes with minimal cardinality, new decision table are mapped. Subsequently, the rules generated from this new table are analyzed for better classification compared to original KDD Cup 99 dataset without prior analysis on reduct and rules generated.

Table 2 is a new description of new decision table for KDD Cup 99 dataset. Original KDD Cup 99 data have 42 attributes including the decision attributes. Nevertheless, the new decision table based on the analyzed reducts reveals 12 condition attributes and 1 decision attribute. These new rules derivation are analyzed based on the rough set benchmark and measurement for better classification than original decision table of KDD Cup 99 dataset.

This section demonstrates the analysis of the generated rule. The rule statistic involve in this analysis are rule support, rule coverage and rule accuracy. Rules computations are done for rule 1 through rule 10. To uncover the most significant rules, these rules are sorted according to their support value. The highest support value is resulted as the most significant rules. All rules are generated with statistics rule. Based on the sorted of highest rule support values, the most significant rule is

*service(ecr_i) AND src_bytes([355, *)) AND dst_bytes([*, 281)) AND hot([*, 1)) AND count([17, *)) AND srv_count([21, *)) AND srv_diff_host_rate(0) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(smurf.)*

This is supported by 4297 for LHS and RHS support value.

Below is the 4 rules of accuracy and coverage with value 1.

service(finger) AND src_bytes([, 239)) AND dst_bytes([*, 281)) AND hot([*, 1)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([*, 80)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([0.03, *)) => type_attack(land.)*

*service(imap4) AND src_bytes([355, *)) AND dst_bytes([1696, *)) AND hot([*, 1)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(0) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([*, 0.01)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(imap.)*

service(http) AND src_bytes([, 239)) AND dst_bytes([1696, *)) AND hot([1, 3)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(1) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([*, 0.01)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(phf.)*

service(telnet) AND src_bytes([239, 355)) AND dst_bytes([281, 1696)) AND hot([1, 3)) AND count([, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([*, 80)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([0.03, *)) => type_attack(loadmodule.)*

Rule Importance Measure (RIM) is used to evaluate the importance of association rules. The analysis of RIM is demonstrated to determine the importance of the attributes. The number of reduct set generated from decision table is 16 and all the attributes of new decision table have the highest RIM percentage which is 100% (refer to Table II). Consequently, all the rules in below attributes are chosen for classification process, thus contributing to better classification accuracy.

No.	Attributes	RIM (%)
1	service	16/16 =100%
2	src_bytes	16/16 =100%
3	dst_bytes	16/16 =100%
4	hot	16/16 =100%
5	count	16/16 =100%
6	srv_count	16/16 =100%
7	srv_diff_host_rate	16/16 =100%
8	dst_host_count	16/16 =100%
9	dst_host_same_srv_rate	16/16 =100%
10	dst_host_diff_srv_rate,	16/16 =100%
11	dst_host_same_src_port_rate	16/16 =100%
12	dst_host_srv_diff_host_rate	16/16 =100%

Table 2. RIM Using Rules from New Decision Table

The significant rules are determined based on;

i) Rule with highest number of support value,

ii) Rule with less length

iii) Rule with accuracy of 1

iv) Rule with coverage of 1

v) Rule with highest number of RIM percentage.

From 1104 rules generated in old decision table, only 5 significant rules are determined. These significant rules are conceded for classification process to improve the classification (refer Table III for complete generated significant rules).

Rules	LHS/ RHS Support	Accuracy	LHS/ RHS Coverage	LHS/ RHS Length
service(ecr_i) AND src_bytes([355, *)) AND dst_bytes([*, 281)) AND hot([*, 1)) AND count([17, *)) AND srv_count([21, *)) AND srv_diff_host_rate(0) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(smurf.)	1,1	1	1	12,1
service(finger) AND src_bytes([*, 239)) AND dst_bytes([*, 281)) AND hot([*, 1)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([*, 80)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([0.03, *)) => type_attack(land.)	1,1	1	1	12,1
service(imap4) AND src_bytes([355, *)) AND dst_bytes([1696, *)) AND hot([*, 1)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(0) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([*, 0.01)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(imap.)	1,1	1	1	12,1
service(http) AND src_bytes([*, 239)) AND dst_bytes([1696, *)) AND hot([1, 3)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(1) AND dst_host_count([255, *)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([*, 0.01)) AND dst_host_srv_diff_host_rate([*, 0.01)) => type_attack(phf.)	4297, 4297	0.2	1	12,1
service(telnet) AND src_bytes([239, 355)) AND dst_bytes([281, 1696)) AND hot([1, 3)) AND count([*, 4)) AND srv_count([*, 5)) AND srv_diff_host_rate(0) AND dst_host_count([*, 80)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate([0.09, *)) AND dst_host_srv_diff_host_rate([0.03, *)) => type_attack(loadmodule.)	1,1	12,1	1	12,1

Table 3. Significant Rules of KDD Cup 99

From previous analysis, it reveals that better classification has been achieved. Generally, IDS do not need to use all the attributes and rules to diagnosis the pattern of attacks, since these will incur long processing time, and no guarantee of better performance. Hence, the core attributes and the significant rule are preferred for quick decision making in determining better result for classification. Table IV illustrates the result of classification performance of original KDD Cup 99 dataset and the new decision table of KDD Cup 99 dataset. Figure 1 illustrates the classification accuracy for original decision table and new decision table of KDD Cup 99 dataset.

Decision Table	**Rule set**	**Overall Accuracy**
Original decision table	All rules	74.3%
New decision table	Selected rules	95.4%

Table 4. Classification Performance between Both Decision Table of KDD Cup 99 Dataset

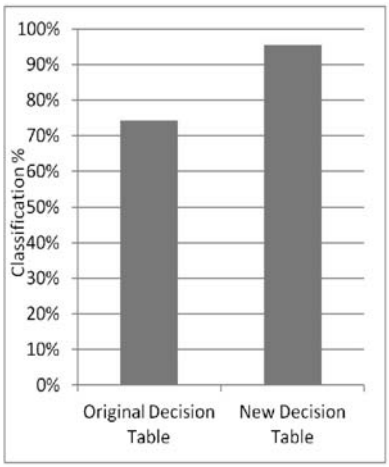

Figure 1. Accuracy for original decision table and new decision table for KDD Cup 99 Dataset

A new generation of decision table for KDD Cup 99 dataset gives a significant impact to the classification rates. Reduct and rules analysis have yielded significant attributes and rules, thus proven to have better classification accuracy compared to the results which employ all attributes, reduct and generated rules.

5. Conclusion

An attempt has been made in this study to explore the significant of reduct and rules that contributing to better classification performance. This study has presented a detail methodology to devise a framework of Rough Set Intrusion Detection System. The process involves a set of procedure principally for reduct generation, rules derivation and classification. Owing to Rosetta flexibility, rough set technique can be applied to the IDS dataset. Several analyses have been achieved to find the significant reduct and rules for better classification. Consequently, the significant attributes are analyzed based on the minimal cardinality and the core attributes of the generated reduct. A new decision table of KDD Cup 99 dataset has been constructed. The rules generated based on this new decision table are analyzed based on highest support value, rules with less length, rules coverage and accuracy with value 1 and Rule Importance Measure (RIM). Nevertheless, the rules with less length cannot establish the significant rules, since the degree of significant depends on high support value for the rules that are being analyzed. As a result, in this study, the influences of using core attributes with minimal cardinality in the course of the generated reduct and significant rules have been examined. An empirical study has been conducted for searching optimal classification. A rough set framework for intrusion detection system dataset is illustrated mutually with an analysis of reduct and derived rules, with entrenchment of their implicit properties for better classification outcomes.

From the experiments and the acquired results, it depicts that Rough set is a remarkable soft computing technique for handling medical data with the existence of missing values. The reduct with core attributes and minimal cardinality assists better classification enhancement. The rules with less length are not efficient as a rule significant measurement. The rules derivation with highest support value, less length and high value or Rule Importance Measure (RIM) are proven to be significant rules in contributing to better classification. The significant of reduct and rules are required to produce better classification result.

6. Acknowledgement

This paper is sponsored by grant RDU 130616. Authors would like to thanks for the supports in making this study a success.

References

[1] Dasarathy, B.V. (2003). Intrusion detection. *Information Fusion*, 4, 243-245.

[2] Yang, L., Jun, L.W., Zhi, H.T., Tian, B.L., Chen, Y. (2009). Building Lightweight Intrusion Detection System using Wrapper-Based Feature Selection Mechanisms. Beijing, China.

[3] Denning, D.E. (1987). An intrusion-detection model. . *IEEE Transactions on Software Engineering*, 13.

[4] Bouzida, Y., Cuppens, F., Cuppens-Boulahia, N., Gombault, S. Efficient Intrusion Detection Using Principal Component Analysis. (2004). *3ème Conférence sur la Sécurité et Architectures Réseaux (SAR)*, La Londe, France

[5] Bose, I. Deciding The Financial Health Of Dot-Coms Using Rough Sets. (2006). *School of Business, University of Hong Kong*.

[6] Pawlak, Z. (1982). Rough Sets. International Journal of Information and Computer Sciences, 11. 341-356.

[7] Pawlak, Z.(1991). Rough Sets In Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers.

[8] Pawlak Z. (1998). Rough Set Theory and its Applications to Data Analysis. *Cybernetics and Systems* 29 (7) 661-688.

[9] Bose, I. Deciding the Financial Health of Dot-Coms Using Rough Sets. (2006). *Information Management.* 43 (7) 835-846.

[10] The KDD99 Dataset. (2010). Retrieved from http://kdd.ics.uci.edu/databases/kddcup99/task.html, retrieved on November 15, 2010.

[11] Jiye L., Cercone N.(2006). Introducing a Rule Important Measure. Transaction on Rough Sets V, Springerlink, 167.