# Feature Selection Using Rough Set in Intrusion Detection

Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia,
81310, Skudai, Johor, Malaysia.
anazida, maarofma and mariyam@fsksm.utm.my

***Abstract*-Most of existing Intrusion Detection Systems use all data features to detect an intrusion. Very little works address the importance of having a small feature subset in designing an efficient intrusion detection system. Some features are redundant and some contribute little to the intrusion detection process. The purpose of this study is to investigate the effectiveness of Rough Set Theory in identifying important features in building an intrusion detection system. Rough Set was also used to classify the data. Here, we used KDD Cup 99 data. Empirical results indicate that Rough Set is comparable to other feature selection techniques deployed by few other researchers.**

## I. Introduction

Intrusion prevention system such as IT policy, firewall and encryption which are widely being practiced are generally inadequate. Despite this rigorous means of filtering and encrypting, attacks still happen and this first layer of defense can still be penetrated. Intrusion Detection Systems (IDS) act as the "second line of defense" and they are placed inside a protected network, looking for known and potential threats in network traffic and/or audit data recorded by hosts [1].

In general, there are two approaches for detecting an intrusion in a computer systems and a computer network: misuse detection and anomaly detection. In misuse detection, an intrusion is detected when the behavior of a system matches with any of the intrusion signatures. Meanwhile the anomaly based IDS will detect an intrusion when the behavior of the system deviates from the normal behavior with certain significant tolerance [2].

IDS can be treated as pattern recognition problem or rather classified as learning system. Thus, an appropriate representation space for learning by selecting relevant attributes to the problem domain is an important issue for learning systems. According to Bello et al.[3], feature selection is useful to reduce dimensionality of training set; it also improves the speed of data manipulation and improves the classification rate by reducing the influence of noise. The goal of feature selection is to find a feature subset maximizing some performance criterion, such as accuracy of classification. Not only that, selecting important features from input data lead to a simplification of the problem, faster and more accurate detection rates. Thus selecting important features is an important issue in intrusion detection [4]. This paper describes an initial work in finding optimal feature subset using Rough Set Theory.

The rest of this paper is structured as follows. Section 2 describes the Rough Set Theory. Section 3 focuses on feature reduction and one exemplary work was discussed. Section 4 describes the data used and experiments carried out. Finally, section 5 concludes the work in this area.

## II. Rough Set Theory

Rough set theory is an extension of conventional set theory that supports approximations in decision making. It is an approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations, which are a classification of the domain of interest into disjoint categories. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset [5]. Rough Set Theory is a mathematical tool for approximate reasoning for decision support and is particularly well suited for classification of objects. It can also be used for feature selection and feature extraction [6].

The main contribution of rough set theory is the concept or reducts. A reduct is a minimal subset of attributes with the same capability of objects classification as the whole set of attributes. Reduct computation of rough set corresponds to feature ranking for IDS. Below is the derivation of how reducts are obtained.

**Definition 1** An information system is defined as a four-tuple as follows, $S$=<$U$, $Q$, $V$, $f$>, where $U=\{x_1, x_2, \ldots, x_n\}$ is a finite set of objects ($n$ is the number of objects); $Q$ is a finite set of attributes, $Q=\{q1, q2, \ldots, qn\}$; $V= \cup_{q \varepsilon Q} V_q$ and $V_q$ is a domain of attribute $q$; $f$:U×Q→$V$ is a total function such that $f(x, q) \in V_q$ for each $q \in Q$, $x \in U$. If the attributes in $S$ can be divided into condition attribute set $C$ and decision attribute set $D$, i.e. $Q=C \cup D$ and $C \cap D=\Phi$, the information system $S$ is called a decision system or decision table.

**Definition 2** Let $IND(P)$, $IND(Q)$ be indiscernible relations determined by attribute sets $P$, $Q$, the $P$ positive region of $Q$, denoted $POS_{IND(P)}(IND(Q))$ is defined as follows:
$POS_{IND(P)}(IND(Q)) = \cup_{X \in U/IND(Q)} / IND(P)\text{-}(X).$

**Definition 3** Let $P$, $Q$, $R$ be an attribute set, we say $R$ is a reduct of $P$ relative to $Q$ if and only if the following conditions are satisfied:

(1) $POS_{IND(R)}\ (IND\ (Q)) = POS_{IND(P)}\ (IND\ (Q))$;
(2) For every $r \in R$ follows that
$POS_{IND(R\text{-}\{r\})}\ (IND\ (Q)) \neq POS_{IND(R)}\ (IND\ (Q))$

Further details can be found in Pawlak [7]. According to Zhang et al. [6], this method produces explainable detection rules and it also has high detection rate for some attacks.

## III. Feature Selection In Intrusion Detection

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features [5].

The work of Zhang et al.[6], exploited the capability of rough set theory in coming up with classification rules in determining the categories of attacks in IDS. Their findings showed that rough set classification attained high detection accuracy (using GA) and the feature ranking was fast. Unfortunately they did not mention the features used for the classification process. Similarly, work of Chebrolu et al. [8] tackled the issue of effectiveness of an IDS in terms of real-time and detection accuracy from the feature reduction perspective. This approach was taken due to large amount of audit data and extraneous features that could complicate the detection process. In their work, features were reduced using two techniques, Bayesian Network (BN) and Classification and Regression Trees (CART). They have experimented using four sets of feature subset which are 12, 17, 19 and all the variables (41) from one network connection. Data used was KDD cup 99. The table below summarizes their findings:

TABLE 1
Performance Of Classifier [8].

| Approach | Variable Size | Type of attack | Accuracy |
|---|---|---|---|
| CART | 12 | Normal | 100% |
| Ensemble | 12 | R2L | 99.29% |
| CART | 17 | Probe | 100% |
| Ensemble | 17 | DoS | 100% |
| CART | 19 | U2R | 84% |

Detail of their work can be found in [8]. Using the same dataset, Sung and Mukkamala [9] ranked six significant features. They used three techniques and compared the performance of these techniques in terms of classification accuracy on the test data. Those techniques were Support Vector Decision Function Ranking (SVDF), Linear Genetic Programming (LGP) and Multivariate Regression Splines (MARS). For detail results, please refer to Ref. [9].

From these reported works, we can conclude that there are features that really significant in classifying the data. Also, it has been proven that there was no single generic classifier that can best classify all the attack types. Instead, in some cases, specific classifier performs better than others. Thus, most of these works on feature selection lead to an ensemble or fusion of multiple classifier IDS.

## IV. Experiment Setup And Results

The data used in this experiment was obtained from database set created by DARPA in the framework of the 1998 Intrusion Detection Evaluation Program (http://www.ll.mit.edu/IST/ideval). In this study, we used the subset that was preprocessed by the Columbia University and distributed as part of the UCI KDD Archive (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html). The labeling of data features as shown in Table 2 is adopted from Chebrolu et al. [8]. The dataset can be classified into five main categories which are *Normal*, *Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R)* and *Probing.* The original data contained 744 MB data with 4,940,000 records. In our experiment, we only used 8000 records where 70% were used for training and another 30% were used for testing. This 70% comprised of 5600 records and the remaining 30% comprised of 2400 records. The dataset contained normal traffic and all categories of attacks.

Here, feature selection was done prior to training. This approach was commonly performed by researchers [8-10]. As mentioned earlier, Chebrolu et al. [8] used 2 feature selection techniques, Bayesian Networks (BN) and Classification and Regression Trees (CART) resulted in ensemble classifier best in classifying R2L and DoS. The former used 12 features and the latter used 17 features respectively. It is an advantage if an IDS can have prior knowledge on the incoming data type so that the right classifier can be initiated. Unfortunately, this is not the case. Thus, it is essential to find an optimum set of features that is generic enough to represent all the data types. Feature selection is an important step in the design and application of any pattern recognizer, including IDS.

TABLE 2
Network Data Feature Labels [8].

| Label | Network Data Features | Label | Network Data Features | Label | Network Data Features | Label | Network Data Features |
|---|---|---|---|---|---|---|---|
| A | duration | L | logged_in | W | count | AH | dst_host_same_srv_rate |
| B | protocol_type | M | num_compromised | X | srv_count | AI | dst_host_diff_srv_rate |
| C | service | N | root_shell | Y | serror_rate | AJ | dst_host_same_src_port_rate |
| D | flag | O | su_attempted | Z | srv_serror_rate | AK | dst_host_srv_diff_host_rate |
| E | src_byte | P | num_root | AA | rerror_rate | AL | dst_host_serror_rate |
| F | dst_bytes | Q | num_file_creations | AB | srv_rerror_rate | AM | dst_host_srv_serror_rate |
| G | land | R | num_shells | AC | same_srv_rate | AN | dst_host_rerror_rate |
| H | wrong_fragment | S | num_access_files | AD | diff_srv_rate | AO | dst_host_srv_rerror_rate |
| I | urgent | T | num_outbound_cmds | AE | srv_diff_host_rate | | |
| J | hot | U | is_host_login | AF | dst_host_count | | |
| K | num_failed_login | V | is_guest_login | AG | dst_host_srv_count | | |

TABLE 3

THE 6 SIGNIFICANT FEATURES OBTAINED BY ROUGH SET

| Technique | Label | Corresponding Features | Description of Features |
|---|---|---|---|
| Rough Set Concept | AO | dst_host_srv_rerror_rate | % of connections from the same host with same service & REJ errors to the destination host during a specified time window |
| | AF | dst_host_count | number of connections from the same host to destination during a specified time window |
| | X | srv_count | number of connections to the same service as the current connection during a specified time window |
| | D | flag | normal or error status of the connection |
| | E | src_byte (source bytes) | number of bytes sent from the host to the destination system |
| | C | service | type of service used to connect (e.g. finger, ftp, telnet, ssh etc.) |

The feature subset obtained will have a great impact on the accuracy of the detection.

As mentioned in earlier section, [9] had ranked the six most significant features using three different techniques namely Support Vector Decision Function (SVDF), Linear Genetic Programming (LGP) and Multivariate Adaptive Regression Splines (MARS). SVDF's proposed features were; B, D, E, W, X and AG. Meanwhile LGP yielded features C, E, L, AA, AE and AI. Finally, MARS suggested features E, X, AA, AG, AH and AI.

Rough Set (RS) reducts obtained using standard Genetic Algorithms were 26 and they were : C, D, E, F, G, J, M, N, P, W, X, Y, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM and AN (refer to Table 2 for features' references). The six most significant features ranked by Rough Set Concept were; C, D, E, X, AF and AO.

Fig. 1, illustrates the comparison between the results obtained using the Rough Set and the findings by Sung and Mukkamala [9] on the 6 most significant features. The description of features' representation obtained using the Rough Set is given in Table 3. Feature E which refers to 'source bytes' (src_bytes), represents the number of bytes sent from the host system to the destination host. Feature E is an important feature as it had constantly been selected by all of the four approaches (Rough Set and three other techniques used by Sung and Mukkamala [9]). It can be observed that 4 features from RS were overlapped with features selected by SVDF, 3 features were overlapped with MARS and 2 features were overlapped with LGP.

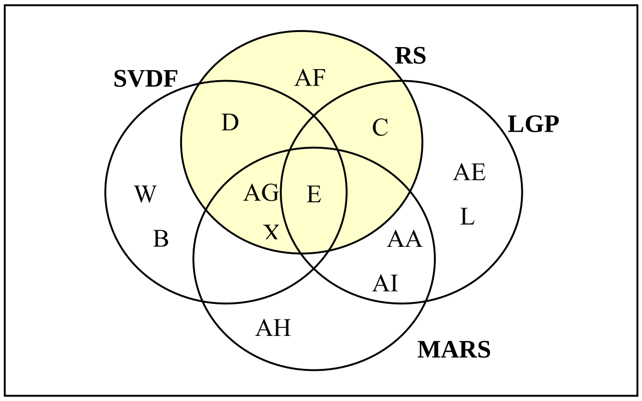

Figure 1. Comparison of 4 feature subsets ranked by various techniques.

Besides its usage in feature selection, we had also applied Rough Set to classify the data to evaluate the performance of 3 feature subsets proposed by [9] and feature subset from Rough Set. The classification results based on feature subsets by MARS, SVM and LGP as proposed by [9] were then compared against classification results based on features obtained using the Rough Set. These results are tabulated in Table 4. LGP seems to superior in classifying data belonging to the normal category, follows by Rough Set. The result on normal also indicates that both SVDF and MARS perform poorly. Low classification rate for normal data is undesirable for an IDS, since it will produce a lot of false alarms.

Meanwhile, the subsequent rows in the table show the classification rate for each attack category. To simplify the analysis, we calculated the mean for all the four attack categories. Mean is important since it generalizes the overall performance of each feature subset when classifying the attack. It is interesting to note that even though LGP performs well in classifying normal data, the opposite is true when classifying the attack. LGP's mean is 89.95% and can be considered very low when compare to others. MARS, SVDF and RS have above 99% classification rate and their performances are at par.

TABLE 4

COMPARISON ON THE CLASSIFICATION ACCURACY USING ROUGH SET AS CLASSIFIER

| Type | | **MARS** | **SVDF** | **LGP** | **RS** |
|---|---|---|---|---|---|
| Normal | | 84.9 | 80.83 | 94.16 | 89.84 |
| Attack | DoS | 99.77 | 99.71 | 99.8 | 99.34 |
| | Probe | 100 | 100 | 100 | 99.63 |
| | U2R | 100 | 100 | 60 | 100 |
| | R2L | 100 | 100 | 100 | 100 |
| | Mean | 99.925 | 99.928 | 89.950 | 99.743 |

RS ranked second after LGP for classifying normal, and perform almost equivalent to MARS and SVDF when classifying attack.

In general, feature subset proposed by RS shows a good performance and comparable to other techniques. This feature subset can be said to be robust since its performance has been almost consistent for all data types. An optimum and robust feature subset is desirable as it contributes to efficient IDS. Efficient in this scope refers to timely and accurate detection.

## V. Conclusion And Future Work

This paper has presented a preprocessing part of an intrusion detection system which is feature selection. An optimum feature subset that can represent data as a whole is essential to the success of an intrusion detection system if both accuracy and speed are to be achieved. The detection process can be expedited if the number of features that are needed to be examined is small. Rough Set has demonstrated its potential capability of selecting an optimum feature subset. The results obtained indicate that the feature subset proposed by Rough Set is robust and has consistent performance through out the experiment. This may be due to Rough Set Theory which heavily relies on the principle of lower and upper approximation and it suits well with the nature of traffic connection that has a grey area between what is normal and intrusive.

We plan to extend the work in terms of accuracy by focusing on fusion of classifiers after a set of optimum feature subset is obtained.

## References

[1] G. Giacinto, F. Roli, and L. Didaci. “Fusion of Multiple Classifiers for Intrusion Detection in Computer Networks”. *Pattern Recognition Letters 24* (2003). Pp. 1795-1803.

[2] Z. Li, and A. Das, (2005). “M of N Features vs. Intrusion Detection.” *ICCSA* 2005, LNCS 3480 Springer Verlag, 2005 pp. 994-1003.

[3] R. Bello, A. Nowe, Y. Caballero, Y. Gomex, , and P. Vrancx, “A Model Based on Ant Colony System and Rough Set Theory to Feature Selection.” *GECCO'05,* June 25-29, 2005, Washington DC, United States. Pp. 275-276.

[4] B. J. Kim and I. K. Kim, “Machine Learning Approach to Realtime Intrusion Detection System.” *In Proceedings of 18th. Australian Join Conference on Artificial Intelligence*, 2005, Sydney, Australia. Vol. 3809. Pp. 153-163.

[5] R. Jensen, and S. Qiang, “Finding Rough Set Reducts with Ant Colony Optimization.” *Proceedings of the 2003 UK Workshop on Computational Intelligence*, 2003, pp. 15-22.

[6] L. Zhang, G. Zhang, L. Yu, J. Zhang, and Y. Bai, “Intrusion Detection Using Rough Set Classification.” *Journal of Zheijiang University Science*. 2004 5(9), pp. 1076-1086.

[7] Z. Pawlak, “Rough Sets: Theoretical Aspects of Reasoning about Data.” Kluwer Academic Publishers, Netherlands. 1991.

[8] S. Chebrolu, A. Abraham, and J. P. Thomas, “Features Deduction and Ensemble Design of Intrusion Detection Systems.” *Journal of Computers and Security*, Volume 24, Issue 4, June 2005, pp. 295-307.

[9] A. H. Sung, and S. Mukkamala, *“The Feature Selection and Intrusion Detection Problems“. Springer Verlag Lecture Notes Computer Science 3321*. 2004, pp. : 468-482.

[10] A. Abraham, and R. Jain, “Soft Computing Models for Network Intrusion Detection Systems”. Classification and Clustering for Knowledge Discovery. Springer Verlag Germany, Chapter 16, 2004, 20 pp.