Rough Set Discretization: Equal Frequency Binning, Entropy/MDL and Semi Naives algorithms of Intrusion Detection System

Noor Suhana Sulaiman
TATi University College
24000 Kemaman, Terengganu, Malaysia
suhana@tatiuc.edu.my

Rohani Abu Bakar
Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300 Kuantan, Pahang, MALAYSIA
rohani@ump.edu.my

ABSTRACT: *Discretization of real value attributes is a vital task in data mining, particularly in the classification problem. Discretization part is also the crucial part resulting the good classification. Empirical results have shown that the quality of classification methods depends on the discretization algorithm in preprocessing step. Universally, discretization is a process of searching for partition of attribute domains into intervals and unifying the values over each interval. Significant discretization technique suit to the Intrusion Detection System (IDS) data need to determine in IDS framework, since IDS data consist of huge records that need to be examined in system. There are many Rough Set discretization technique that can be used, among of them are Semi Naives and Equal Frequency Binning.*

Keywords: Rough Set, Semi Naives, Equal Frequency Binning, Discretization, Intrusion Detection System

Received: 10 May 2017, Revised 7 June 2017, Accepted 14 June 2017

1. Introduction

Large amount of audit data that an IDS needs to examine and analysis is difficult because extraneous features can make it harder to detect suspicious behavior patterns [1]. Audit data capture various features of the connections. For example, the audit data would show the source and destination bytes of a TCP connection, or the number of failed login attempts or duration of a connection. In the network intrusion detection, the system needs to handle massive amounts of network data in real-time manner, typically, the network data contains a large number of features [2], which significantly increases the load of IDS, but at the same time, there are many irrelevant and redundant features that will decline detection accuracy during the intrusion detection process based on machine learning mechanism and bring additional of the complexity of learning algorithms. All of these require IDS must be able to select the right subset of the most important features to improve the detection accuracy and efficiency.

The question in the large number of features that can be monitored for IDS, which are truly useful, which are less significant, and which may be useless are relevant because the elimination of useless features (or audit trail reduction) enhances the accuracy of detection while speeding up the computation, thus improving the overall performance of the detection mechanism [3].

Rough set theory has been successfully applied in many fields such as machine learning, pattern recognition, decision support and data mining. However, rough set can't deal with continuous attributes. One solution to this problem is to partition numeric variables into a number of intervals and treat each interval as a category. This process of partitioning continuous variables into categories is usually termed discretization. Discretization is the calculation of a core for discrete attribute dataset, containing strongly relevant attributes, and reducts, contains a core plus additional weakly relevant attributes, such that each reduct is satisfactory to determine concepts in the dataset. Based on a set of reducts for a dataset, some criteria for attribute selection can be formed which a reduct containing minimal is set of attributes [4]. Discrete values are intervals of number which are more concise to represent and specify, easier to use and comprehend as they are closer to a knowledge-level representation than continuous values. Discretization technique involves searching for cuts that determine intervals. All values that lie within each interval are mapped to the same value, in effect converting numerical attributes that treated as being symbolic. The search for cuts is performed on the internal integer representation of the input decision system. Due to the critical scenario on IDS data, significant discretization technique suit to IDS data need to analyze.

Many analyses have been done to come out with good discretization technique. The authors in [5] introduced the discretization algorithm with division of equidistance, and another discretization algorithm with division of Equal frequency, literature [6] proposed a kind of discretization algorithm combined polynomial hyperplanes and support vector machine (SVM), a discretization algorithm is proposed based on the theory of cloud in literature [7], Li et al. [8] proposed a discretization algorithm based on hierarchy clustering, literature [9] provided a discretization algorithm based on the division with interval. Above these algorithms were solved to some problems for decision table discretization, but could engender new incompatible problem for object in the decision table. literature [10] proofed that the calculation optimal breakpoint set is NP-hard problem, therefore, most of discretization algorithms are belong to heuristic algorithm, literature [11] provided discretization algorithm and greedy algorithm based on importance of the breakpoint, this two kinds of algorithm belong to global discretization algorithms and have high recognition rate are recognized by researchers. Literature [12] proposed discretization algorithm based on the attribute importance, which belong to local discretization algorithms, literature [13] introduced discretization algorithm based on the information entropy. Above those methods consider the incompatibility of decision table and many of them can obtain good results, but the complexity of the algorithm is high.

Not all the discretization algorithms are significant to IDS data for better classification. Hence, significant discretization technique must be chosen to contribute better classification. The results show that chosen method give better classification performance compared to standard rough set in terms of accuracy.

2. Rough Set Discretization Theory

Rough-Set Theory (RST) was introduced by Polish logician, Professor Zdzis³aw Pawlak in 1982 to cope with imprecise or vague concepts. 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 [14,15]. 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 [16].

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.

In an approximation space, it is required that a set of equivalence relations is given first, and it also contains one kind of definitely formal method, which maps any subset of equivalence relations into an equivalent relation on the universe. The concept of approximation space gives a kind of approximation classification analytica1 methods when processing the total classification of objects belong a certain universe, under insufficient information.

The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest. Given an information system $A = (U, A)$, *let* $X \subseteq U$ be a set of objects and $B \subseteq A$ be a selected set of attributes. B-lower approximation $\underline{B}X$ is;

$$\underline{B}X = \{x \in U : [x]B \subseteq X\}$$

The lower approximation is the set of objects which can certainly be classified into decision X. These objects must have their equivalence class totally contained in the set X. This means that in a lower approximation, all objects can be discerned from those outside the set $\underline{B}X$ [17].

The upper approximation is a description of the domain objects which contains all objects which possibly belong to the concept.

$$\overline{B}X = \{x \in U : [x], \cap X \neq 0$$

The upper approximation of those objects that can possibly be classified is represented into X. These objects are indiscernible from one or more elements in X. The upper approximation is a superset of the lower approximation and X.

3. Intrusion Detection System dataset

This study employed KDD Cup (1999) dataset which is an IDS benchmark dataset, prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Lab. This dataset has been used for the Third International Knowledge Discovery and Data Mining Tools Competition. The task of this competition was to build a network detector to find "bad" connections and "good" connections. Each record has 41 attributes describing different features and a label assigned to each either as an attack type or as normal. The protocols that are considered in KDD dataset are TCP, UDP, and ICMP. Transmission Control Protocol (TCP) is an important protocol of the Internet Protocol Suite at the Transport Layer which is the fourth layer of the OSI model. It is a reliable connection-oriented protocol which implies that data sent from one side is sure to reach the destination in the same order. TCP splits the data into labeled packets and sends them across the network. TCP is used for many protocols such as HTTP and Email Transfer. User Datagram Protocol (UDP) is similar in behavior to TCP except that it is unreliable and connection-less protocol. As the data travels over unreliable media, the data may not reach in the same order, packets may be missing and duplication of packets is possible. This protocol is a transaction-oriented protocol which is useful in situations where delivery of data in certain time is more important than loosing few packets over the network. It is useful in situations where error checking and correction is possible in application level. Internet Control Message Protocol (ICMP) is basically used for communication between two connected computers. The main purpose of ICMP is to send messages over networked computers. The ICMP redirect the messages and it is used by routers to provide the up-to-date routing information to hosts, which initially have minimal routing information. When a host receives an ICMP redirect message, it will modify its routing table according to the message.

There are 38 numeric features and 3 symbolic features, falling into the following four categories: [18]

- Basic features: 9 basic features were used to describe each individual TCP Connection, basic features to every network connection like duration of connection, service requested, and bytes transferred between source and destination machine etc.
- Content features: 13 domain knowledge related features were used to indicate suspicious behavior having no sequential patterns in the network traffic, llike logged in flag, number of failed logins, hot indicators, etc.
- Time-based traffic features: 9 features were used to summarize the connections in the past 2 s that had the same destination host or the same service as the current connection. Time-based traffic were collected by observing various connections in "two-second" time window with respect to current connection, such as SYN error rates, Rejection rates, number of different services requested etc
- Host-based traffic features: 10 features were constructed using a window of 100 connections to the same host instead of a time window, because slow scan attacks may occupy a much larger time interval than 2 s., based on the past 100 connections similar to the one under consideration.

The original data contains 744MB data with 4,940,000 records. Each record was marked with a value of classification attribute,

which only had two values: normal record or DDoS attack There were 9,775 records in training data with 1,928 normal records and 7,847 DoS attacks. DoS records had one of five types: Neptune, *smurf*, pod, teardrop and back attack. There were 10,952 records in testing data with 2,251 normal records and 6,184 DoS records respectively. Five new attack types just appeared in tested data not in training data. There were totally 334 of such new attacks: *mailbomb, land, proceetable, warezmaster* and apache attacks. These new attacks were used to find weather the system could detect new attacks.

4. Description of Rough Set Discretization

A discretization [19] replaces value sets of conditional real-valued attributes with intervals. The replacement ensures that a consistent decision system is obtained which assuming a given consistent decision system by substituting original values of objects in the decision table by the unique names of the intervals comprising these values. This substantially reduces the size of the value sets of real-valued attributes. Discrete values have important roles in data mining and knowledge discovery. Rules with discretize values are normally shorter and more understandable and discretization can improve accuracy.

A decision table is composed of a 4-tuple as follows:

$S = < U, Q \cup \{d\}, V, f >$, where

U: a finite set of N objects $\{x_1, x_2, ...x_N\}$:

Q: Finite set of n condition attributes $\{q_1, q_2, ...,q_n\}$ (a nonempty set), and d is decision attribute;

$V = U_{q \in Q} V_q$ wher V_q is a domain of the attribute q.

$f : U \times Q \cup d \rightarrow V$ is the total decision function called information function such that $f(x, q) \in V_q$ for every $q \in Q \cup d, x \in U$. The decision table can be represented as a finite data table, in which the columns are labeled by attributes, the rows by objects and the entry in column q_j and row x_i has the value $f(x_i, q_j)$. Each row in the table describes the information about some objects in S.

Assume $V_q = [l_q, r_q) \subset R$, where R the set of real numbers is, and assume that S is consistent decision table [1]. The following notion and description about discretization is referred to reference [20].

Definition 1 Any pair (q,c) , where $q \in Q$ and $c \in R$,defines a partition of V_q into left-hand-side and right hand-side interval. The pair (q, c) is called a cut on V_q.

For an attributeis $q \in Q$, $D_q = \{(q,c_1^q), (q, c_2^q), ...,(q, c_{kq}^q)\}$ is composed by all the cuts, where $k_q \in N$, and $l_q = c_0^q < c_1^q < c_2^q < ... < c_{kq}^q < c_{kq+1}^q) = r_q$, defines a partition on V_q into subintervals i.e. $V_q = [c_0^q, c_1^q) \cup [c_1^q, c_2^q) \cup ...\cup [c_{kq}^q, c^q_{kq+1})$. Hence, any set of cuts on condition attributes $D = \cup D_a$ transforms the original decision table S into discrete decision table $S^D =< U, Q \cup \{d\}, V^D, f^D >$, where $f^D(x, q) = i \Leftrightarrow f(x, q) \in [c_i^q, c_{i+1}^q)$, and $x \in U, i \in \{0, 1, ...,k_q\}, q \in Q$.

After discretization, the original decision table is replaced with the new one. And different sets of cuts will construct different new decision table. It is obvious that discretization process is associated with loss of information. Usually the task of discretization is to determine a minimal set of cuts from a given decision table and keeping the discernibility. The selected cuts can be evaluated by the following criteria:

- Consistency of D. For any objects $u, v \in U$, they are satisfying if u ,v are discerned by Q, then u, v are discerned by D;
- Irreducibility. There is no $D' \subset D$, satisfying the consistency;
- Optimality. For any D′satisfying consistency, it follows card (D) $\leq$ card (D′), then D is optimal cuts.

The algorithms of sample data discretization include Equal Distance Binning, Equal Frequency Binning (EFB), Naive algorithm, Semi Naive algorithm, Boolean Reasoning algorithm and Entropy algorithm [21]. In [22], Semi Naive algorithm is applied and the Equal Distance Binning algorithm is employed in [23]. Since the Boolean Reasoning algorithm is good at finding minimal subsets of cuts that preserve the original discernibility in decision table, it was chosen to carry out the discretization at first, and then

those attributes which do not need to keep the discernibility can be discretized by equal frequency algorithm.

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 significant discretization algorithms are analyzed suits to IDS data. The classification is implemented using standard voting classifier.

Training data is discretized using EFB, Semi Naives and Entropy/MDL algorithm 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 Equal Frequency Binning algorithm, there are 17 bins allocated to the value of numeric numbers of IDS data, refer to Table 1 below.

No of bin	Value of bin	No of bin	Value of bin
1	[0.5, 3.5)	9	[0.5, 1.5)
2	[238.5, 354.5)	10	[0.5, 1.5)
3	[280.5, 1695.5)	11	[0.5, 1.5)
4	[0.5, 1.5)	12	[3.5, 16.5)
5	[0.5, 2.5)	13	[4.5, 20.5)
6	[0.5, 1.5)	14	[79.5, 254.5)
7	[0.5, 1.5)	15	[254.5)
8	[0.5, 4.5)	16	[0.5, 8.5)
		17	[0.5, 2.5)

Table 1. Value of bin by Equal Frequency Binning (EFB)

This section demonstrates the analysis of the generated rule. Based on the sorted of highest rule support values, the rule with the highest value, derived from Equal Frequency Binning is

duration([, 1)) AND service(ecr_i) AND src_bytes([355, *)) AND dst_bytes([*, 281)) AND hot([*, 1)) AND logged_in(0) 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)) AND dst_host_srv_rerror_rate(0) => type_attack(smurf.)*

This is supported by 4297 for LHS and RHS support value, with accuracy of 1. The number of generated rules that need to be analyze are 186,32 rules.

The generated rule with the highest support value, derived from Entropy/MDL is

service(ecr_i) AND src_bytes([1032, 1033)) AND dst_bytes([, 5)) AND dst_host_diff_srv_rate(0) AND dst_host_same_src_port_rate(1.00) => type_attack(smurf.)*

This is supported by 4288 for LHS and RHS support value, with accuracy of 1. The number of generated rules that need to be process are 99,476 rules.

The generated rule with the highest support value, derived from Semi Naives is

service(http) AND flag(SF) AND src_bytes([101, 612)) AND num_failed_logins(0) AND num_shells([, 2)) AND num_access_files(0) AND count([*, 86)) AND srv_rerror_rate(0) AND same_srv_rate([1, *)) AND dst_host_same_srv_rate(1) AND dst_host_diff_srv_rate(0) AND dst_host_srv_diff_host_rate([*, 0.47)) => type_attack(normal.)*

This is supported by 12508 for LHS and RHS support value. There are 147,485 number of rules generated to be analyze into the classification process.

From previous analysis, it reveals that better classification has been achieved. Generally, the significant discretization algorithm is important in resulting the better percentage of classification suit to IDS data. Table 2 illustrates the result of classification performance of KDD Cup 99 data by using three different discretization algoritms which are EFB, Entropy/MDL and Semi Naives. Figure 1 illustrates the result of classification percentage.

Discretization Algorithm	**Overall Accuracy**
Equal Frequency Binning	99.86%
Entropy/MDL	75.07%
Semi Naives	74.79%

Table 2. Classification Performance of 3 Discretization Algorithm to IDS Dataset

Figure 1. Classification percentage of KDD Cup 99 Dataset by using 3 different discretization algorithm

Equal Frequency Binning is the most significant discretization algorithm suit to IDS data, consequently gives a significant impact to the classification rates. The new outcome yield that different discretization algorithm probing different number of generated rules to be examine, thus there is a different types of significant rule in same IDS data.

5. Conclusion

The main conclusions drawn from the computational experiments performed so far are the following. The significant discretization algorithm, thus proven to have better classification accuracy compared to the results which employ all attributes, reduct and generated rules.

An attempt has been made in this study to explore the significant of discretization technique suit to IDS data. The process involves a set of procedure principally for eliminate missing value n redundant features, discretization process, generation of reduct and rules and classification process to Rosetta flexibility, rough set technique can be applied to the IDS dataset. Several analyses have been achieved to probe the significant discretization algorithm for better classification. Thus, significant discretization algorithm is well probe suit to IDS data. As a result, in this study, the influences of using significant discretization

algorithm in the course of IDS classification have been examined. An empirical study has been conducted for searching optimalclassification. It shows different types of discretization algorithm generate different number of rules to be examine, thus IDS classification have different type of significant rule to be process. The significant discretization algorithm and the significantrule are preferred for quick decision making in determining better result of classification.

References

[1] Lee, W., Stolfo, S., & Mok, K. A Data Mining Framework for Building Intrusion Detection Models. (1999). *Proceedings of the IEEE Symposium on Security and Privacy.*

[2] Xiang, C., Bing-Xiang, L., & Yi-Lai, Z. (2010). Attribute Reduction Method Applied to IDS. Information engineering Institute,Jingdezhen Ceramic Institute.

[3] Nguyen, H.S. (1998). Discretization Problem for Rough Sets Methods. *Proc. of First Intern. Conf. on Rough Sets and Current Trend in Computing (RSCTC'98)*, Warsaw, Poland, 545-552.

[4] Wang Guoyin. (2001). Rough set theory and knowledge acquisition [M]. Xi 'an jiaotong university press

[5] He Yaqun,Hu Shousong. (2003) A New M ethod for Continuous Value Attribute Discretization in Rough Set Theory [J]. Journal of Nanjing University of Aeronautics & Astronautics, 35(3): 213-215.

[6] Li Xingsheng. (2003). A decision system based on cloud model Discretization method [J]. Pattern Recognition and Artificial Intelligence, 16(3):33-38.

[7] M X Li, C D Wu, Z H Han. (2004). A Hierarchical Clustering Method For Attribute Discretization in Rough Set Theory[C]. Proceedings of The third International Conference on Machine Learning and Cybernetics : 3650-3654.

[8] C T Su, J H Hsu. (2005). An Extended Chi2 Algorithm for Discretization of Real Value Attributes[J]. IEEE Transaction on Knowledge and Data Engineering, 17(3): 437-441.

[9] A Skowron, C Rauszer. (1991). The Discernibility Functions Matrics And Functions in Information Systems[C]. Intelligent Decision Support: 331-362.

[10] H S Nguyen. (2006). Approximate Boolean Reasoning: Foundations And Applications in Aata Mining[J]. Transaction on Rough Set : 334-506.

[11] Hou Lijuan, Wang Guoyin, Nie Neng. (2000). discretization problem of Rough set theory [J]. Computer science, 27(12): 89-94.

[12] Xie Hong, Cheng Haozhong, Niu Dongxiao. (2005). Discretization Algorithm based on the information entropy of rough set [J]. Journal of computer, 28(9): 1570-1574.

[13] 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.

[14] Pawlak, Z. (1991). Rough Sets, Theoretical Aspects of Reasoning about Data, *Kluwer Academic*, Dordrecht.

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

[16] Guoyong, W. (2001). Rough Sets Theory And Knowledge Acquisition. *Xi'an: Xi'an Jiaotong University Press.*

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

[18] Pawlak, Z., & Skowron, A. (2007). Rough sets and Boolean reasoning. Institute of Mathematics, Warsaw University, ul. Banacha 2, 02-097 Warsaw, Poland Information Sciences 177; 2007. 41–73.

[19] Jian-Hua Dai,Yuan-Xiang Li. (2002) Study on discretization based on rough set theory.Proceedings of the First International Conference on Machine Learning and Cybernetics,Beijng

[20] G. Y. Wang. (2001). Rough theory and knowledge acquirement. Xian Jiao Tong press.

[21] X.R. Wang, R.S. Xu, W.Q. Zhang. (2004). Network Intrusion Detection System Based on Rough Set Theory. ComputerScience. Vol.3INo.11, pp. 80-82

[22] L. H. Zhang, G. H. Zhang et.al. (2004). Intrusion detection using rough set classification. Journal of Zhejiang University SCIENCE. Vol.5 No.9, pp:1076-1086.