Evaluation of Rough Set Theory Based Network Traffic Data Classifier Using Different Discretization Method

Nandita Sengupta, *Member, IACSIT* and Jaya Sil, *Member, IEEE*

***Abstract*—In information security, intrusion detection is a challenging task for which designing of an efficient classifier is most important. In the paper, network traffic data is classified using rough set theory where discretization of data is a necessary preprocessing step. Different discretization methods are available and selection of one has great impact on classification accuracy, time complexity and system adaptability. Three discretization methods are applied on continuous KDD network data namely, rough set exploration system (RSES), supervised and unsupervised discretization methods to evaluate the classifier accuracy. It has been observed that supervised discretization yields best accuracy for rough set classification and provides system adaptability.**

***Index Terms*—Classification, cuts, discretization, network traffic, rough set theory.**

I. Introduction

Online classification of network traffic data is very important to develop intrusion detection system (IDS) that automatically monitors the flow of network packets. Existing works on intrusion detection have been carried out to classify the network traffic as anomaly or normal. A majority of current IDS follow signature based approach [1] in which similar to virus scanners, events are detected that match specific predefined patterns known as "signatures". The limitation of these signature based IDS is their failure to identify novel attacks and even minor variation of patterns are not detected accurately. In addition, sometimes IDS generate false alarm for alerting network administrator due to failure of handling imprecise data which has high possibility to appear in network traffic data. Therefore, accuracy, computation time and system learning are the key issues to be addressed properly for classifying such data.

Classification is an important task in data mining research that facilitates analysis of huge amount of data. Rough Set Theory (RST) [2] is based on mathematical concept can handle vagueness in classification of data. However, prior to applying RST, the data is discretized and selection of discretization procedure has great impact on classification accuracy. In the paper, network traffic data [3] of KDD has been considered for generating training and testing patterns. In order to apply RST, the datasets are discretized using RSES , supervised and unsupervised based techniques. After discretization [4], using indiscernibility relation of RST, a minimum subset of attributes of the dataset is selected, called reducts by applying exhaustive algorithm [5]. Rules are generated from the reducts and classifiers are built using rule set classifier [6]. Finally, classification accuracy has been expressed in form of confusion matrix [7], which provides information about actual and predicted classification achieved by a classification system. Classification accuracy is compared based on the discretized tables generated from cuts using RSES software, supervised technique and unsupervised technique using WEKA software.

Section II describes about rough set theory, section III mentions about discretization, section IV depicts experimental results and section V concludes the paper and mentions future work.

Manuscript received February 20, 2012; revised April 26, 2012.

N. Sengupta is with University College of Bahrain, P O Box 55040, Manama, Bahrain (e-mail: ngupta@ucb.edu.bh).

J. Sil is with Bengal Engineering and Science University Shibpur, P O Botanic Garden, Howrah, West Bengal, Pin 711103 (e-mail: js@cs.becs.ac.in).

II. Rough Set Theory

A. Information System

Information system [8] is nothing but Data table. Here we consider U as an nonempty set of objects, a data table is a tuple (U,A,Va)a€A , where A is a set of attributes a:U→Va and Va is a set of values for the attribute a. The set of attributes can be divided into two subsets, conditional set of attributes, C and decision set of attributes, D. C and D both are subsets of A, $C \subset A$ and D=A-C, Conditional set of attributes represent all the features or attributes of objects and decision set of attributes represent the classification of objects.

B. Set Approximation

Equivalence classes are called indiscernible [8]. If the values of conditional attributes of some objects are same, those objects are declared as indiscernible. In a data table where indiscernibility relations are found, table can be defined in two ways, consistent and inconsistent. If all the objects in indiscernibility relations are classified in the same class, the table is called as consistent on the other hand, if all the objects in indiscernibility relations are not classified as the same class, the table is said as inconsistent. In that case some features/attributes may have not been reflected precisely in the data table. Rough set is defined (see, fig. 1) as a pair of crisp sets, lower approximation $\underline{C}X$ and upper approximation $\overline{C}X$, where X is the target set. Lower approximation set is known as positive region because X is characterized by a particular decision value. There are also indiscernibility classes which contain only some tuples in X, which cannot be classified exactly. These are the objects in boundary region and

mathematically, represented as $\overline{C}X - \underline{C}X$. The elements which are in U- $\overline{C}X$, belongs to the negative region. X is crisp or precise when $\underline{C}X = \overline{C}X$, means boundary region is empty.

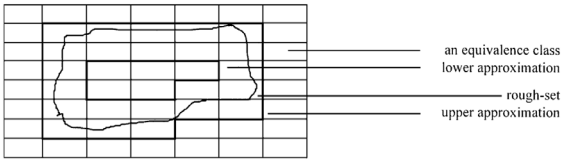

Fig.1. Rough Set with Lower and Upper Approximation

C. Reducts and Rule Generation

In order to reduce redundant and insignificant attributes, concept of reducts is emerged in RST, a subset of conditional attributes representing the whole data table. Finding reduct is NP hard problem and many researchers [9] are working on fixing up algorithm for finding reduct. Decision rules [10] are generated from reducts and used for classification of objects.

III. Discretization

Discretization is a process in which continuous attributes are divided into some intervals and represented by new values according to the intervals. In machine learning, discretization is an important approach for handling continuous attributes. There are many methods available [11] for discretization methods, equal width discretization (EWD), equal frequency discretization (EFD), fuzzy discretization (FD), entropy minimization discretization (EMD), iterative discretization (ID), proportional *k*-interval discretization (PKID), lazy discretization (LD), nondisjoint discretization (NDD), weighted proportional *k*-interval discretization (WPKID). Different discretization method is effective for different machine learning system. In the paper, supervised and unsupervised discretization methods are applied to discretize the continuous attributes. Supervised method considers class information while unsupervised discretization method does not consider class information intensively.

IV. Experimental Results

TABLE I: Comparison of Classification Accuracy

	Accuracy		
	Discretization using RSES	Supervised Discretizatio n using WEKA	Unsupervised Discretization using WEKA
Anomaly	0.994	0.991	0.927
Normal	0.944	0.979	0.947
Total Accuracy	0.986	0.989	0.930

In the paper, subset of KDD network traffic data has been considered where the actual dataset contains 11850 objects and each object has 42 attributes. Our information system consists of 5332 objects for training and 1185 objects for testing the classifier. Three discretization methods, namely RSES, supervised and unsupervised methods are applied and then the discretized dataset is classified using RST after generating reducts and rule sets. Result for comparison of accuracy is shown in Table I.

A. Discretization through Cut Generation Using RSES

Cuts are defined as the partitions of a range for any continuous attribute. Cuts are calculated satisfying some natural conditions in the information system. Discretization of real valued attributes is done from the cut only. In the first case, whole data set ('KDDTest-21') is split into two tables, containing 1185 and 10665 objects respectively. Then again the second table is split into two tables, having 5332 and 5333 objects respectively. Using the cut set of table ('KDDTest-21'_0.9_0.5_1_C), the discretized tables 'KDDTest-21'_0.9_0.5_1D and 'KDDTest-21'_0.1D are formed. Decision rules are generated as shown in the table, 'KDDTest_21'_0.9_0.5_1D_R and finally, the table 'KDDTest-21'_0.1D_F is constructed after classification. Fig. 2 represents the action flow graph of the above case. Classification result is represented in the form of confusion matrix, shown in Table II.

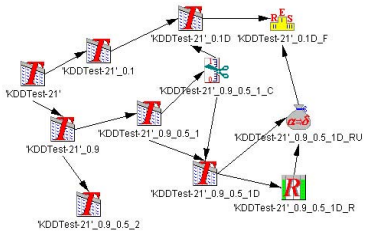

Fig. 2. Rough flowchart of actions of classification using RSES method of discretization

TABLE II: Confusion Matrix using RSES Method of Discretization

		Predicted				
		anomaly	normal	Obj[a]	Acc[b]	Cov[c]
Actual	anomaly	798	5	962	.994	.835
	normal	8	134	223	.944	.637
	True positive rate	0.99	0.96			

a. Obj refers No. of Objects, b. Acc refers Accuracy, c. Cov refers Coversge

B. Supervised Discretization using WEKA

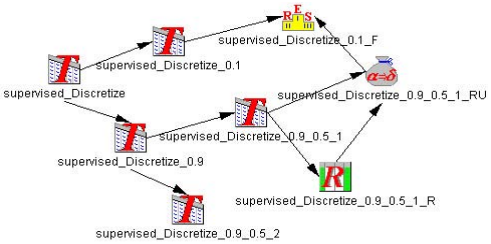

Fig. 3. Flowchart of actions of classification using supervised method of discretization

Whole dataset, 'KDDTest-21' is considered for supervised discretization using WEKA software. Supervised discretized table named as "supervised_Discretize", which is imported in RSES environment. Tables of cut, discretization, rule generation and finally classification results are obtained like the earlier case and fig. 3 represents the action flow graph for

this case. Classification table is "supervised_Discretize_0.1_F" is represented in table III.

TABLE III: CONFUSION MATRIX USING USING SUPERVISED DISCRETIZATION

		Predicted				
		anomaly	*normal*	*Obj*[a]	*Acc*[b]	*Cov*[c]
Actual	anomaly	801	7	971	.991	.832
	normal	3	138	214	.979	.659
	True positive rate	1	0.95			

a. Obj refers No. of Objects, b. Acc refers Accuracy, c. Cov refers Coversge

C. *Unsupervised Discretization using WEKA*

Whole dataset, 'KDDTest-21' is considered for unsupervised discretization in WEKA software. Unsupervised discretized table named as "unsupervised_Discretize" Fig. 4 represents the action flow graph for this case considering the steps of earlier two cases. Classification table is "unsupervised_Discretize_0.1_F" represented in table IV.

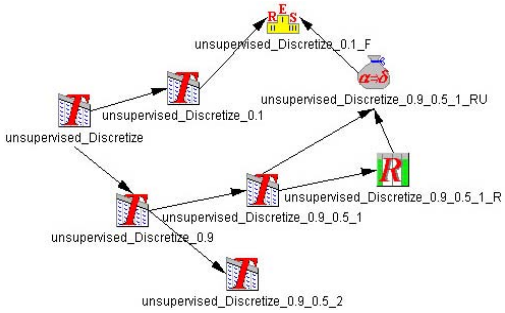

Fig. 4. Flowchart of actions of classification using Unsupervised Method of Discretization

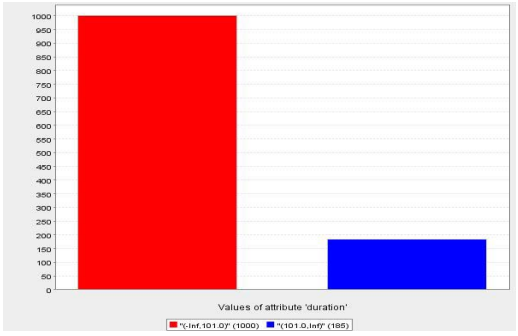

Fig. 5. Range of values of "Duration" attribute after discretization using RSES

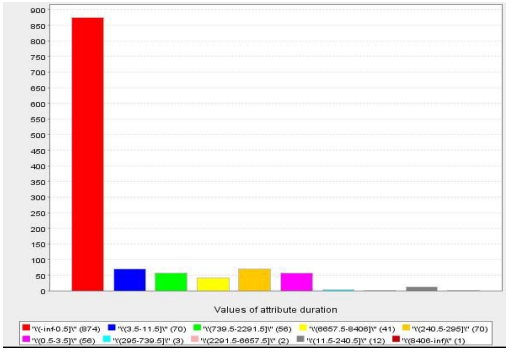

Fig .6. Range of values of "Duration" attribute after supervised discretization

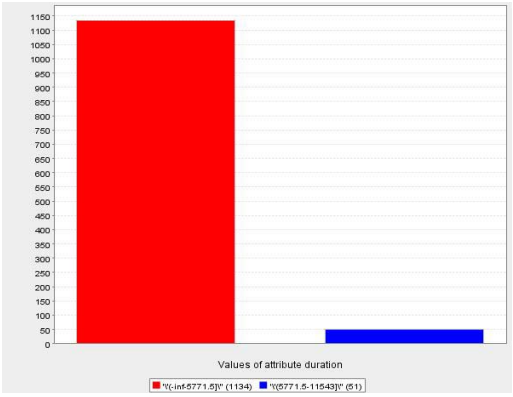

Fig. 7. Range of values of "Duration" attribute after unsupervised discretization

TABLE IV: CONFUSION MATRIX USING UNSUPERVISED DISCRETIZATION

		Predicted				
		anomaly	normal	Obj[a]	Acc[b]	Cov[c]
Actual	anomaly	707	56	993	.927	.768
	normal	7	126	192	.947	.693
	True positive rate	0.99	0.96			

a. Obj refers No. of Objects, b. Acc refers Accuracy, c. Cov refers Coversge

V. CONCLUSION AND FUTURE WORK

Rough set theory based classification of network traffic data handles minimal set of attributes and vagueness that reduces complexity of the IDS. Comparison of the classification result in the three cases is demonstrated and the second case, i.e., classification after supervised discretization yields the best average accuracy. Fig. 5, 6 and 7 are provided to show range of values of only one attribute "duration" after discretization, in three different methods. Comparing these figures, it can be decided that supervised discretization is much more rigorous that the other two methods. In the future work, optimized rule sets will be generated to decrease computational time of classification.

REFERENCES

[1] S. Neelakantan and S. Rao, "A threat-aware signature based intrusion-detection approach for obtaining network-specific useful alarms," in *Proc. The Third International Conference on Internet Monitoring and Protection*. 2008.

[2] T. Beaubouef and F. E. Petry, "Uncertainty modeling for database design using intuitionistic and rough set theory," *Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology*, vol. 20, no. 3, 2009.

[3] "Nsl-kdd data set for network-based intrusion detection systems."Available : http://iscx.ca/NSL-KDD/

[4] P. Blajdo, J. W. Grzymala-Busse, Z. S. Hippe, M. Knap, T. Mroczek, and L. Piatek, "A comparison of six approaches to discretization—a rough set perspective," *Rough Sets and Knowledge Technology*, Lecture Notes in Computer Science, vol. 5009, pp. 31-38, 2008,

[5] L. Gaojun and Z. Yan, "Credit assessment of contractors: a rough set method," Tsinghua Science and Technology, vol. 11, no. 3, 2006.

[6] S. Kumar, S. Atri, and H. L. Mandoria, "A Combined Classifier to Detect Landmines Using Rough Set Theory and Hebb Net Learning and Fuzzy Filter as Neural Networks," in *Proc. ICSPS, 2009.*

[7] N. Sengupta and J. Sil, "Decision making system for network traffic*,"* in *Proc. KBIE Jan. 2011.*

[8] N. Sengupta and J. Sil, "An integrated approach to information retrieval using RST, FS and SOM," in *Proc. ICIS2008*, Bahrain.2008.

[9] J. Zhang, J. Wang, D. Li, H. He, and J. Sun, "A new heuristic reduct algorithm base on rough sets theory," *Advances in Web-Age Information Management*, Lecture Notes in Computer Science, Vol. 2762, 247-253, 2003,
[10] K. Dembczyński, R. Pindur, and R. Susmaga, "Generation of Exhaustive Set of Rules within Dominance-based Rough Set Approach," in *Proc. International Workshop on Rough Sets in Knowledge Discovery and Soft Computing*, 2003.
[11] Y. Yang and G. I. Webb, "A comparative study of discretization methods for Naive-Bayes classifiers," in *Proc. PKAW*, 2002.

Nandita Sengupta has done her Bachelor and Master of Engineering from Bengal Engineering and Science University, Shibpur, India. She has 21 years of working experience and last 10 years with academics. Currently, she is in University College of Bahrain, Bahrain. Her area of interest is Analysis of Algorithm, Theory of Computation, and Network Computing.

Jaya Sil received her B.E. in Electronics and Telecommunication Engineering and M.E. degree in Computer Science and Engineering, from Bengal Engineering College and Jadavpur University, India in 1984 and 1986, respectively. She received her Ph.D. degree in Engineering (Computer Science and Engineering) from Jadavpur University, India in 1996. She has 24 years teaching experience and presently acts as Professor of Computer Science and Technology Department of Bengal Engineering and Science University, Shibpur, India. Her research areas Artificial Intelligence and Soft Computing, Image Processing, and Bio-Informatics. Dr. Sil received fellowship for post doctoral research and visited Nanyang Institute of Technology, Singapore on 2002-2003. She has visited DKFZ Lab in Heidelberg, Germany and Tsinghua University, Bejing, China for collaborative research.