

Comparison of Different Rule Calculation Method for Rough Set Theory

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

Abstract—In the paper, information table for network traffic data is discretized by supervised learning using WEKA and from this discretized data, three different algorithms for calculation of rules are applied, genetic algorithm, covering algorithm and LEM2 algorithm. Rough set classification through three different table rule set are done and accuracies are observed. It has been concluded that out of these three methods of rule calculation, classification through table rule set using covering algorithm yields best accuracy. In fourth table of classification, rules have been calculated through reduct generation but in that case, rough set classification produces same total accuracy of classification through covering algorithm. RSES software is used for rule set calculation, reduct generation and rough set classification.

Index Terms—Classification, network traffic, reducts, rough set theory, rule generation.

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, discretization is done through supervised learning using WEKA. After discretization, rules

are generated in four different ways. In first three cases, rules are generated using genetic algorithm [4], covering algorithm [5] and LEM2 algorithm [6] respectively. Using indiscernibility relation of RST, a minimum subset of attributes of the dataset is selected, called reducts by applying exhaustive algorithm [7]. Then in the fourth case, rules are generated from reducts. Classifiers are built using rule set classifier [8]. Finally, classification accuracy has been expressed in form of confusion matrix [9], which provides information about actual and predicted classification achieved by a classification system. Classification accuracy is compared based on the four sets of rules. It has been observed that for direct rule generation, out of those three algorithms, covering algorithm produces best ‘total accuracy’ and rule from reduct also yields the same ‘total accuracy’ as achieved from covering algorithm.

Section II describes about rough set theory, section III mentions about algorithms for rule generation, section IV depicts experimental results and section V concludes the paper.

II. ROUGH SET THEORY

A. Information System

Information system [10] is nothing but Data table. Here we consider U as an nonempty set of objects, a data table is a tuple $(U, A, Va)_{a \in A}$, where A is a set of attributes $a: U \rightarrow Va$ 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. 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 [11] are working on fixing up algorithm for finding reduct. Decision rules [12] are generated from reducts and used for classification of objects.

III. ALGORITHMS FOR RULE GENERATION

Three different algorithms, genetic algorithm, covering algorithm and LEM2 algorithm are applied for three different cases of rule generation.

Manuscript received February 20, 2012; revised April 29, 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).

IV. EXPERIMENTAL RESULTS

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 2133 objects for training and 1185 objects for testing the classifier. Three different algorithm methods for rule generation are applied. Then rough set classifier is applied from those three different rule sets. In the fourth case, reducts are generated by using exhaustive algorithm and then rules are generated from the reducts. Result for comparison of accuracy is shown in Table I.

TABLE I: COMPARISON OF CLASSIFICATION ACCURACY

	Accuracy			
	Rules using Genetic Algorithm	Rules using Covering Algorithm	Rules using LEM2 Algorithm	Rules through Reduct
Anomaly	0.981	0.999	0.993	0.992
Normal	0.915	0.939	0.911	0.983
Total Accuracy	0.968	0.991	0.977	0.991

Continuous attributes in whole data set are discretized using supervised learning algorithm using WEKA software. Whole data set ('supervised_Discretized') is split into two tables, containing 1185 and 10665 objects respectively. Then again the second table is split into two tables, having 2133 and 8532 objects respectively. Table with 2133 no. of objects have been considered as training table. From this training set, four sets of rules are generated and from those rule sets rough set classifier has been used to classify the traffic.

In the first case, genetic algorithm is used to get the set of rules and then classification result has been found out in form of confusion matrix which is mentioned in Table II.

TABLE II :CONFUSION MATRIX USING GENETIC ALGORITHM FOR RULE GENERATION

	GENERATION					
	Predicted					
		anomaly	normal	Obj ^a	Acc ^b	Cov ^c
Actual	anomaly	933	18	951	.981	1
	normal	20	214	234	.915	1
	True positive rate	0.98	0.92			

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

In the second case, covering algorithm is used to get the set of rules and then classification result has been found out in form of confusion matrix which is mentioned in table III.

TABLE III :CONFUSION MATRIX USING COVERING ALGORITHM FOR RULE GENERATION

	GENERATION					
	Predicted					Observed
		anomaly	normal	Obj ^a	Acc ^b	Cov ^c
Actual	anomaly	697	1	951	.999	.734
	normal	6	93	234	.939	.423
	True positive rate	0.99	0.99			

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

In the third case, LEM2 algorithm is used to get the set of rules and then classification result has been found out in form of confusion matrix which is mentioned in Table IV.

TABLE IV: CONFUSION MATRIX USING LEM2 ALGORITHM FOR RULE GENERATION

GENERATION						
	Predicted					
		anomaly	normal	Obj ^a	Acc ^b	Cov ^c
Act ual	anomaly	825	6	951	.993	.874
	normal	17	173	234	.911	.812
	True positive rate	0.98	0.97			

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

In the fourth case, first reduces are generated by using exhaustive algorithm and rules are generated from those reduces. Classification result has been found out in form of confusion matrix which is mentioned in Table V.

TABLE V: CONFUSION MATRIX USING RULES GENERATED FROM REDUCTS

	Predicted					
		anomaly	normal	Obj ^a	Acc ^b	Cov ^c
Actual	anomaly	748	6	951	.992	.793
	normal	2	117	234	.983	.509
	True positive rate	1	0.95			

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

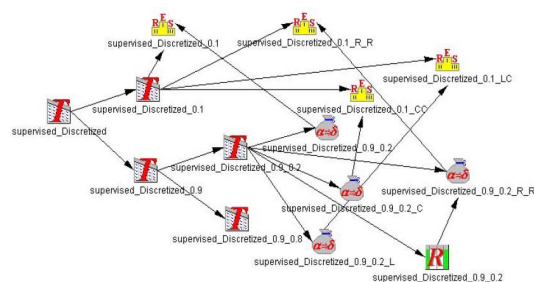


Fig. 1. Flowchart of actions of classification using different rule generation methods

Fig. 1 represents the action flow graph of the above case. Length of rule in all above four cases have been produced in Table VI.

TABLE VI : COMPARISON OF RULE LENGTH

	Accuracy			
	Rules using Genetic Algorithm	Rules using Covering Algorithm	Rules using LEM2 Algorithm	Rules through Reduct
Minimal	1	1	24	11
Maximal	9	1	41	11
Mean	3	1	36	11

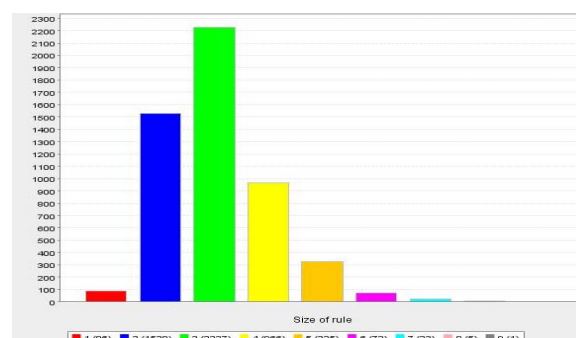


Fig. 2. Rule length for rule set generated using genetic algorithm

Fig. 2, 3, 4 and 5 are provided to show rule lengths of rule sets of four different cases.

V. CONCLUSION

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 four cases is demonstrated and the second case, i.e., classification through rule generation using covering algorithm yields the best average accuracy and it is also observed that second case and fourth case, where rules are generated through reduct, produces same average accuracy result.



Fig. 3. Rule length for rule set generated using covering algorithm

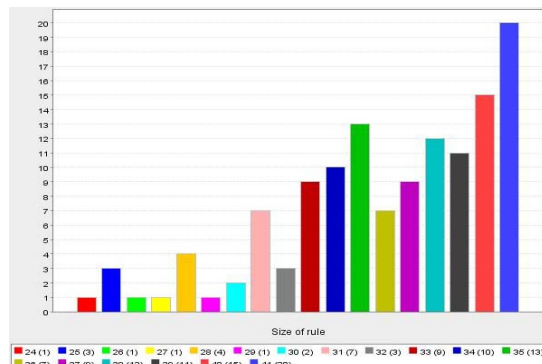


Fig. 4. Rule length for rule set generated using LEM2 algorithm

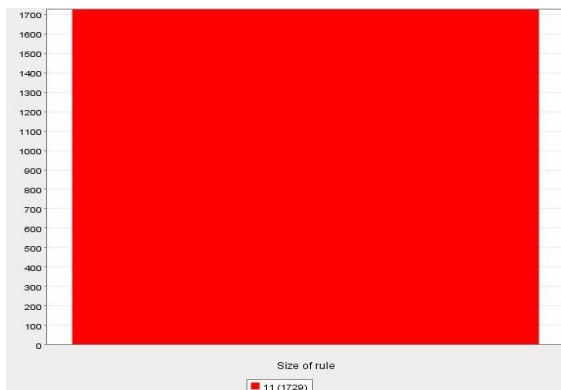


Fig. 5. Rule length of rule set generated from reduct

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 & Fuzzy Systems: Applications in Engineering and Technology*, vol. 20, 2009.
- [3] "Nsl-kdd data set for network-based intrusion detection systems." Available : <http://iscx.ca/NSL-KDD/>
- [4] P. P. W. Waiswa and V. Baryamureeba, "Extraction of interesting association rules using genetic algorithms," *International Journal of Computing and ICT Research*, vol. 2, no. 1, 2008.
- [5] J. Błaszczyński, R. Słowiński, and M. Szeląg, "Sequential covering rule induction algorithm for variable consistency rough set approaches," *Information Sciences: an International Journal*, vol. 181, 2011.
- [6] R. Decker and F. Kroll, "Classification in marketing research by means of LEM2 generated rules," R. Decker, and H.-J. Lenz (Eds.), *Advances in Data Analysis*, Berlin: Springer, pp. 425 – 432, 2007.
- [7] L. Gaojun, and Z. Yan, "Credit assessment of contractors: a rough set method," *Tsinghua Science and Technology*, vol. 11, no. 3, 2006.
- [8] S. Kumar, S. Atri, and H. L. Mandoria, "A Combined Classifier to Detect Landmines Using Rough Set Theory and Hebb Net Learning & Fuzzy Filter as Neural Networks," in *Proc. ICSPS, 2009*.
- [9] N. Sengupta and J. Sil, "Decision making system for network traffic," in *Proc. KBIE Jan*, 2011.
- [10] N. Sengupta and J. Sil, "An integrated approach to information retrieval using RST, FS and SOM," in *Proc. ICIS2008*, Bahrain.
- [11] 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*, 2003.
- [12] 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.



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, Beijing, China for collaborative research.