Journal of Theoretical and Applied Information Technology
30th June 2015. Vol.76. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING ROUGH SET THEORY ATTRIBUTES SELECTION OF KDD CUP 1999

[1]HAMID H. JEBUR [2]MOHD AIZAINI MAAROF [3]ANAZIDA ZAINAL
Universiti Teknologi Malaysia, Faculty of Computing
81310 Skudai, Johor, Malaysia
E-mail: [1]hamedhh59@yahoo.com , [2]Aizaini@utm.my, [3]Anazida@utm.my

ABSTRACT

Attribute selection (Feature Selection) is a significant technique for data preprocessing and dimensionality reduction**.** Rough set has been used for attribute selection with great success. The optimal solution of rough set attribute selection is a subset of attributes called a reduct. Rough set uses approximation during reduction process to handle information inconsistency. However, some rough set approaches to attribute selection are inadequate at finding optimal reductions as no perfect heuristic can ensure optimality. Applying rough set for selecting the optimal subset of KDD Cup 1999 does not guarantee finding the optimal reduct of each class of this dataset due to the overlap between the lower and upper approximation of each class and the overlap between the reducts of all classes. This paper introduces a new approach to enhance the reduct of all classes by overcoming the overlap problem of rough set through adding union and voting attributes of all dataset classes as new reducts in addition to the normal reduct. The all reducts were evaluated by using different classification algorithms. The approach led to generate two generic attributes sets that achieved high and comparable accuracy rates as the normal attributes of rough set for the same dataset.

Keywords: Rough set; Attribute selection; Reduct; approximations; KDD Cup 1999

1. INTRODUCTION

Zdzisław Pawlak introduced the Rough Set Theory (RST) in 1982 as a mathematical method to deal with vague and fuzzy information [1, 2, 3]. Rough set theory is commonly used for attributes selection in information system [4]. The attribute selection optimal solution is an attributes subset with minimal number of attributes called a reduct [5]. The rough set produces all potential reducts and chooses the one with minimum cardinality and highest dependency [6]. The rough set theory main concept is an indiscernibility relation between two objects or more that have similar values with a considered subset of attributes [1]. Inconsistent information in rough set theory is handled using two approximations; the upper and lower approximations [1]. The lower includes all objects that definitely belong to the set, while the upper approximation contains all objects that may belong to the set. The difference between the two approximations represents the rough set boundary region [7, 8, 9]. The approximations and the boundary region of a dataset B is illustrated in figure1.

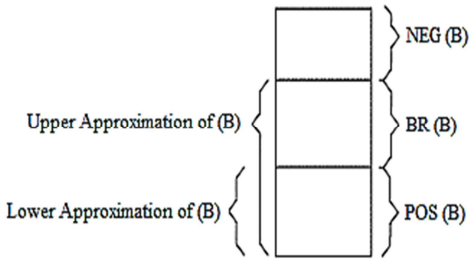

Figure 1: Rough Set Approximations [10]

POS (B) = lower approximation (definitely member of B)
NEG (B) = not member of B
BR (B) = upper approximation (potentially member of B)

Nevertheless, reducts calculation is a major problem in RS theory [11, 12, 13] and calculating the optimal reduct is an NP-hard problem [3, 13, 4]. The reduct contains the significant attributes and irrelevant attributes [3].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Anyway, some rough set methods to attribute selection are incapable to find optimal reduct [2, 9, 12]. As an example, Johnson's algorithm is implemented in the Rosetta toolkit for reduct in rough set. The algorithm is a simple greedy single reduct algorithm that considers the recurrent attributes in every repetition as the most related to the reduct. Yet, it does not guarantee to find the best solution, even though it finds a solution closer to the optimum in some cases [5].

The literature provides many rough set algorithms for attributes selection, but most of them are inefficient to fine optimal reduct [8, 14]. Applying rough set for selecting the optimal subset of KDD Cup 1999 does not ensure finding the optimal reduct of each class of these data due to two reasons. First, there is an overlap between the attributes of upper and lower approximations of each class through the boundary region. Second, there may be an overlap between the reducts of all classes. To overcome this problem, this paper introduces a new approach to enhance the reducts of all classes while keeping high accuracy rate. The approach involves adding union and voting attributes of all data classes as new reducts. The paper is organized as follows. Section 2 presented the related work. Sec 3 shows the research data. The proposed approach is introduced in section 4. The results and discussion are explained in section 5, while conclusion is presented in section 6.

2. RELATED WORK

Researchers utilized different algorithms and approaches to enhance the rough set theory reducts including heuristic and meta-heuristic algorithms [14]. Wang et al [2] proposed rough sets and particle swarm optimization (PSO) to overcome the inadequacy of hill-climbing rough set approaches in attribute selection. Wang et al [3] proposed a rough set approach based on scatter search algorithm for attribute selection. Mafarja and Eleyan [6] proposed an attributes reduction method based on Ant Colony algorithm and rough set theory. Arafat et al. [15] proposed Ant Colony algorithm to optimize the reduction of rough set theory. Hedar et al. [11] suggested a Tabu search to enhance attribute reduction in rough set theory. Jabbar and Zainudin [16] used the water cycle algorithm (WCA) to determine the minimum reduction in the rough set theory. Mafarja and Abdullah [17] proposed the record-to-record travel optimization algorithm and rough set for attribute reduction. Most of previous researches are more computational cost and time consuming, which are overlooked in many times if the researches achieve better results. However, in spite of much research, no heuristic could assure optimality [2].

3. RESEARCH DATA

DARPA dataset is still used as a benchmark for testing most of the intrusion detection system [18, 19, 20]. KDD Cup 1999 includes (41) attributes shown in table 1. The dataset is classified into five main categories, which are Normal, Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing. A subset of 10% of KDD Cup 1999 dataset preprocessed by [21, 22] is used in this study. The training data includes 5,092 records, and testing data includes 6,890 records. These records have same distribution of KDD Cup 1999 dataset. The motivation for using KDD Cup1999 is that many algorithms and approaches were used for attributes selection using this dataset, but they selected different types and number of attributes [21, 27, 28].

Table 1: KDD CUP 1999 Attributes [22]

4. THE PROPOSED APPROACH

To prove the research objective, Johnson's algorithm is selected to demonstrate the research approach. Johnson's algorithm rough set is implemented on the total attributes of KDD Cup 1999 to select the minimal subset of attributes for each class. Johnson's algorithm generates single subset of attributes called reduct. This subset is considered in this research as normal attributes to differentiate it from other subsets of attributes. Johnson's algorithm rough set is based on selecting the frequent attributes as the reduct core. However, this algorithm does not guarantee to find the best reduct. The proposed approach utilizes the function of the algorithm on the normal attributes of each class since the reduct of each class is considered inefficient. The approach performs its function to solve the overlap problem in two ways. First, it selects the union attributes of all dataset classes to ensure that the reduct includes all the relevant attributes for each class. Second, it selects the most frequent attributes in all classes as the core

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attributes for a generic reduct for all classes. The frequent attributes are considered as voting attributes. The approach idea can be explained as follows. The reduct of each class contains essential attributes and redundant ones and may be miss some attributes found in the other classes. Therefore, Union attributes are selected because they inevitably contain all the relevant attributes of each class. Due to the high probability of the existence of redundant attributes, the most frequent attributes in all categories are selected and considered essential and necessary attributes of each class. Therefore, two set of generic attribute are produced. These sets are evaluated by using classification algorithms to compare their performances. The approach is illustrated in figure 2, and the pseudo code is listed below.

Pseudo Code

1. Apply RST algorithm to each class of KDD Cup 1999
2. Define normal attributers set of each class
3. Find union attributes set of all classes normal attributes
4. Apply voting technique to find voting attributes set
   *For all normal attributes of all class:
   *Find the repetition of each attributes
   *If the number of attribute repetition is ≥ 3 adds the attribute to the voting attribute set
   *Define the voting attribute set
5. Apply classification algorithm to evaluate normal, union and voting attributes sets
6. The set with high accuracy = final attributes set

Figure 2: The Proposed Approach

5. RESULTS AND DISCUSSION

ROSETTA GUI version 1.4.40 toolkit is used to implement Johnson's algorithm rough set for attribute selection. Rosetta is an open source collection of C++ classes and routines used for data mining and machine learning in general and particularly for rough set theory [23]. Weka 3.7.9 is used for classification. Weka is open source software developed by Waikato University in New Zealand. It includes several machine-learning algorithms for data mining. The results of the RST for each class (normal attributes) and for the union and voting attributes for all classes are shown in tables 2 and 3.

Table 2: Normal Attributes of Each Class

Class	Normal Attributes									
Class 1	2	3	6	10	23	32	33			
Class 2	2	3	23	27	32	33	36			
Class 3	2	3	6	7	10	23	25	32	33	36
Class 4	2	6	22	23	27	33	36			
Class 5	2	6	7	10	12	23	32	33		

Table3: Union and Voting Attributes of All lass

Union Attributes	2	3	6	7	10	12	22	23	25	27	32	33	36
Voting Attributes	2	3	6		10			23			32	33	36

It can be shown from tables 2&3 that the normal attributes of all classes include various sets. Class 1, 2 and 4 have (7) different attributes, while classes 3 and 5 have (10) and (8) attributes respectively. Union attributes number equals to 13, and voting attributes number equals to 8. The three sets (normal, union and voting) of attributes are evaluated using various algorithms in Weka namely J48, Random Forest and K Star. The J48 algorithm is a version of C4.5 decision tree for classification [24]. The Random Forests uses the random selection of training data samples to make a set of un-pruned regression trees [25]. The K Star is an instance-based classifier that uses an entropy-based distance function on the contrary of other instance-based learners. The K Star classifies an instance by comparing it with pre-classified examples [26]. A full training dataset of each

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dataset class is used to train the classifiers to build the classification models that were evaluated on the test data of the same classes. The classification results are shown in tables 4,5,6,7 and 8.

Table 4: Classification Results of Class1

Class 1			
Classification Algorithm	Normal Attributes	Voting Attributes	Union Attributes
J48	99.6429	99.6429	99.5714
K Star	98.0714	98.0714	98.2857
Random Forest	99.6429	99.5	99.6429

Table 4 indicates that three sets of attributes achieve almost the same accuracy rates with slight differences. Voting attributes achieve the same accuracy of normal ones when using J48 and K Star algorithms, while union attributes achieve higher accuracy when using K Star algorithm.

Table 5: Classification Results of Class2

Class 2			
ClassificationAlgorithm	Normal Attributes	Voting Attributes	Union Attributes
J48	100	100	99.7143
K Star	99.7143	99.7143	99.7143
Random Forest	100	100	99.9286

Table 5 also indicates slight differences among the three sets of attributes; however, voting attributes achieve better accuracy when using all algorithms.

Table 6: Classification Results of Class3

Class 3			
Classification Algorithm	Normal Attributes	Voting Attributes	Union Attributes
J48	99.9048	99.9048	99.6667
K Star	96.3333	96.3333	96.5714
Random Forest	99.9524	99.7619	100

Class 3 classification results as shown in table 6 indicate that voting attributes achieve same accuracy when using J48 and K Star algorithms, while union attributes achieve better accuracy when using K Star and Random Forest.

Table 7: Classification Results of Class4

Class 4			
Classification Algorithm	Normal Attributes	Voting Attributes	Union Attributes
J48	100	100	100
K Star	100	100	100
Random Forest	100	100	100

Table 7 shows a high performance for all algorithms, where union and voting attributes achieve maximum accuracy same as normal attributes

Table 8: Classification Results of Class5

Class 5			
Classification Algorithm	Normal Attributes	Voting Attributes	Union Attributes
J48	99.9526	99.9684	99.921
K Star	99.8736	99.6365	99.8894
Random Forest	99.9842	99.9526	99.9684

Table 8 shows a convergence between the results of the three attributes sets and the algorithms. Voting attributes achieve better accuracy when using J48 algorithm, while union attributes achieve better accuracy when using K Star algorithm

Generally, the union and voting attributes achieve high performance close to a large extent to the normal attributes and sometimes surpass them. For all classes, J48 algorithm achieves high accuracy rate equals to the normal attributes (sometimes better) with voting attributes. On the other hand, K Star algorithm achieves same results as J48 algorithm for all classes but with union attributes. The slight differences between the results of the algorithms may be due to the effect of algorithms behaviors and structures.

6. CONCLUSION

This paper introduces a new approach to enhance the RST reduct for KDD Cup 1999 dataset by overcoming the overlap problem between the reduct of RST of all dataset classes through adding new reducts represented by union and voting attributes. The proposed approach

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

yielded two subsets of generic attributes of KDD Cup 1999 that include (8) and (13) attributes. Both sets are evaluated using different algorithms and revealed high and comparable performance comparing with normal attributes of the rough set reduct. The motivation for this research is the inefficiency of some rough set approaches to find the minimal reduct and the vagueness between the rough set approximations. The approach achieves two important issues; the first is providing generic attributes of KDD Cup 1999. The second one is providing a method to mitigate the overlapping problem of rough set. The future works aims to apply the same approach to several reducts of same algorithm for the same dataset and with different datasets.

ACKNOWLEDGEMENT

This research is supported by UTM research university grant (PY/2014/02479).

REFERENCES

[1] Crossingham, B. Using optimization techniques to granulize rough set partitions (2009)

[2] Wang, X., Yang, J., Teng, X., Xia, W., & Jensen, R. Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters, 28(4), 459-471(2007)

[3] Wang, J., Zhang, Q., Abdel-Rahman, H., & Abdel-Monem, M. I. A rough set approach to feature selection based on scatter search meta-heuristic. Journal of Systems Science and Complexity, 27(1), 157-168 (2014)

[4] Nazri, M. Z. A., Ayob M., Yassen E. T. Evolutionary Iterated Local Search for Rough Set Feature Reduction Problems. International Journal of Advancements in Computing Technology (IJACT), Volume 5(10) (2013)

[5] Li, X. Attribute Selection Methods in Rough Set Theory. Master thesis, San José State University (2014)

[6]Mafarja, M., & Eleyan, D. Ant Colony Optimization based Feature Selection in Rough Set Theory. International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 1(2) (2013)

[7] Rissino, S., & Lambert-Torres, G. Rough Set Theory–Fundamental Concepts, Principals, Data Extraction, and Applications. Data Mining and Knowledge Discovery in Real Life Applications, J. Ponce and A. Karahoca (Eds.), pp. 438, InTech Publishers, Vienna, Austria (2009)

[8] Pawlak, Z. Rough set theory and its applications. Journal of telecommunications and information technology, 7-10 (2002)

[9] Jensen, R., & Shen, Q. Finding rough set reducts with ant colony optimization. In Proceedings of the 2003 UK workshop on computational intelligence (Vol. 1, No. 2) (2003)

[10] Lambert-Torres, G.; Rossi, R.; Jardini, J.A.; Alves da Silva, A.P. & Quintana, V.H. Power System Security Analysis based on Rough Classification, In: Rough Fuzzy Hybridization: A New Trend in Decision-Making, Pal, S. & Skowron, A. (Ed.), pp. 263-300, Springer-Verlag Co., ISBN 981-4021-00-8, Secaucus-USA (1999)

[11] Hedar, A. R., Wang, J., & Fukushima, M. Tabu search for attribute reduction in rough set theory. Soft Computing, 12(9), 909-918 (2008)

[12] Sun, L., Xu, J. C., Xue, Z. A., & Zhang, L. J. Rough entropy-based feature selection and its application. Journal of Information and Computational Science, 8(9), 1525-1532 (2011)

[13] Novaković, J., STRBAC, P., & Bulatović, D. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043, 21(1) (2011)

[14] Anaraki, J. R., & Eftekhari, M. Rough set based feature selection: A Review. In Information and Knowledge Technology (IKT), 2013 5th Conference on (pp. 301-306). IEEE (2013)

[15] Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. Using rough set and ant colony optimization in feature selection. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2(1) (2013)

[16] Jabbar, A., & Zainudin, S. Water Cycle Algorithm for Attribute Reduction Problems in Rough Set Theory. Journal of Theoretical & Applied Information Technology, 61(1) (2014)

[17] Mafarja, M., & Abdullah, S.. Record-to-record travel algorithm for attribute reduction in rough set theory. Journal of Theoretical and Applied Information Technology, 49(2) (2013)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18]Mahoney, M. V.. Network traffic anomaly detection based on packet bytes. Proceedings of the 2003 ACM symposium on Applied computing, 2003: ACM, 346-350 (2003)

[19]Aggarwal M., Amrita. Performance Analysis of Different Feature Selection Methods in Intrusion Detection. International Journal of Scientific & Technology Research Vol. 2(6) (2013)

[20]Porto-Díaz, I., Martínez-Rego, D., Alonso-Betanzos, A., & Fontenla-Romero, O. Combining feature selection and local modelling in the kdd cup 99 dataset. In Artificial Neural Networks–ICANN 2009 (pp. 824-833). Springer Berlin Heidelberg (2009)

[21] Zainal, A., Maarof, M. A., & Shamsuddin, S. M. Feature selection using rough set in intrusion detection. In TENCON 2006. 2006 IEEE Region 10 Conference (pp. 1-4). IEEE (2006)

[22]Chebrolu, S., Abraham, A., & Thomas, J. P. Feature deduction and ensemble design of intrusion detection systems. Computers & Security, 24(4), 295-307 (2005)

[23]Wang, H., & Wang, S. Rough Set Analysis for Total Information Quality Management. In ASAC,Vol. 30 (7) (2009)

[24]Patil, T. R., & Sherekar, M. S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Computer Science and Applications, 6(2) (2013)

[25]Ali, J., Khan, R., Ahmad, N., & Maqsood, I. Random forests and decision trees. IJCSI International Journal of Computer Science Issues, 9(5) (2012)

[26]Mahmood, D. Y., & Hussein, M. A. Intrusion Detection System Based on K-Star Classifier and Feature Set Reduction. International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE) Vol. 15, 107-112 (2013)

[27]Olusola, A. A., Oladele, A. S., & Abosede, D. O. Analysis of KDD'99 Intrusion detection dataset for selection of relevance features. In Proceedings of the World Congress on Engineering and Computer Science Vol.1, 20-22 (2010)

[28]Chen, R. C., Cheng, K. F., Chen, Y. H., & Hsieh, C. F. Using rough set and support vector machine for network intrusion detection system. In Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference (pp. 465-470). IEEE (2009)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: KDD Cup 1999 Attributes [22]

No	Network attributes	No	Network attributes	No	Network attributes	No	Network attributes
1	duration	12	logged_in	23	count	34	dst_host_same_srv_rate
2	protocol_type	13	num_compromised	24	srv_count	35	dst_host_diff_srv_rate
3	service	14	root shell	25	serror_rate	36	dst_host_same_src_port_rate
4	flag	15	su_ attemped	26	srv_serror_rate	37	dst_host_srv_diff_host_rate
5	src_bytes	16	Num_root	27	rerror_rate	38	dst_host_serror_rate
6	dst_bytes	17	Num_file_creations	28	srv_rerror_rate	39	dst_host_srv_serror_rate
7	land	18	Num_shells	29	same_srv_rate	40	dst_host_rerror_rate
8	wrong fragment	19	num_access_files	30	diff_srv_rate	41	dst_host_srv_rerror_rate
9	urgent	20	num_outbound_cmds	31	srv_diff_host_rate		
10	hot	21	is_host_login	32	dst_host_count		
11	Num_failed_logins	22	is_guest_login	33	dst_host_srv_count		

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

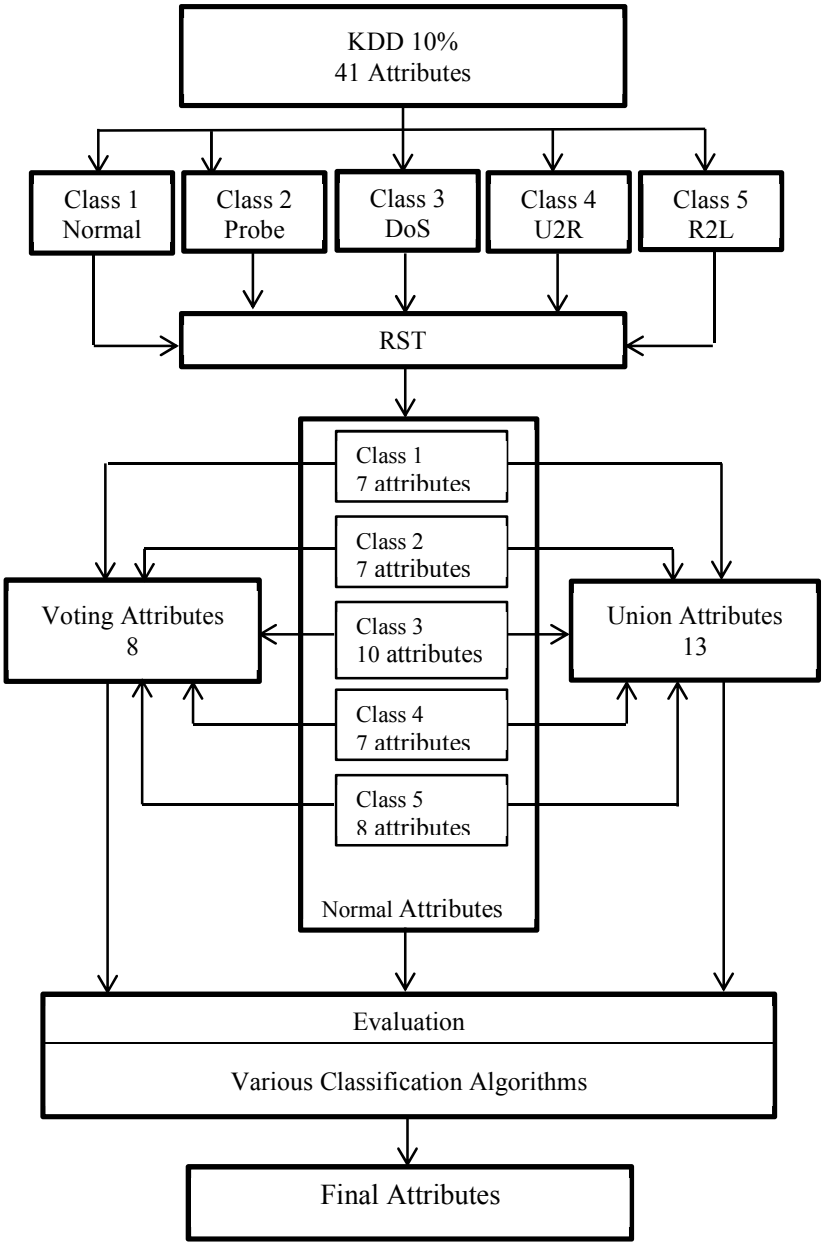

Figure 2: The Proposed Approach