International Journal of Information Processing, 6(4), 64-74, 2012
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Compact Reduct Formation for Classification Rule Set Generation using Rough Set Theory

Asit Kr. Das[a], Shampa Sengupta[b]

[a]Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howraha 711 103, West Bengal, India. Contact: akdas@cs.becs.ac.in

[b]Department of Information Technology, MCKV Institute of Engineering, Liluah, Howraha 711 204, West Bengal, India. Contact: shampa2512@yahoo.co.in

Large volume of data is collected frequently in the course of daily work in different fields. Typically, the datasets constantly grow accumulating a large number of features, most of which are not relevant in decision-making. Moreover, the information often lacks completeness and has relatively low information density. Dimensionality reduction is a fundamental area of research in data mining domain. Rough Set Theory (RST), based on a mathematical concept, has become very popular in dimensionality reduction of large datasets. The method is used to determine a subset of attributes called *'reduct'* which can predict the decision concepts. In the paper, the concepts of discernibility relation and attribute dependency are integrated for the formation of compact reduct set and the concept of decision matrix is used for generation of classification rules, which not only reduces the complexity but also helps to achieve higher accuracy of the decision system. A sample decision system is used as an example for generation of compact reduct set and formation of classification rules. The proposed method has been applied on wine dataset collected from the UCI repository and the classification accuracy is calculated for all possible reducts by some existing classifiers. Using decision matrix approach classification rules are also generated from the reduct and 92% of classification accuracy is achieved. Experimental result shows the efficiency of the proposed method.

Keywords: Attribute Dependency, Core, Classification Rules, Decision Matrix, Decision System, Discernibility Relation, Reduct.

1. INTRODUCTION

Rough Set Theory (RST) is an efficient mathematical concept used for dimensionality reduction [1] [2] as well as classification of data [3] [4]. A series of reduction algorithms [5] [6] were constructed for all kinds of applications based on rough set models. However, determining minimal set of attributes, called reduct, is NP-complete [7] problem. There is usually more than one reduct for real world datasets. It is not very clear which subset of reducts should be selected for learning. Exhaustive search for finding reduct is infeasible and therefore, heuristic methods based on distinct measures of significance of attributes, such as discernibility matrix [8] based algorithm, dependency based [9] algorithm, mutual information [10] based algorithm, genetic algorithm [11] and dynamic reduction algorithm [12] are applied. In reality, there are multiple reducts in a given information system used for developing classifiers, amongst which the best performer is chosen as the final solution to the problem. But this is not always true and according to the Occam's razor and minimal description length principle [13]-[15], the minimal reduct is preferred. However, Roman [16] has found that the minimal reduct is good for ideal situations where a given dataset fully represents a domain of interest. But for real life situations and limited size datasets, those other than the minimal reducts might be better for prediction. Selecting a reduct with good performance is time expensive, as there might be many reducts of a given dataset. Therefore, obtaining a best per-

Table 7
Reducts of Wine Dataset with Accuracies Given by Various Classifiers

Classifiers	Reducts					Average accuracy for all reducts
	{BFGJKM}	{BGHJKM}	{ADEGJL}	{ABEGJKL}	{ADEFGJL}	
Naïve Bayes	97.18	98.87	99.44	99.44	98.31	98.65
SVM	95.48	96.61	97.18	97.74	96.61	96.72
SMO	95.48	94.92	96.05	96.61	96.06	95.82
KSTAR	96.05	98.31	94.92	94.92	94.92	95.82
Bagging	94.35	94.92	96.05	94.92	95.48	95.14
MultiClass	98.87	99.44	98.87	98.31	97.18	98.53
J48	97.74	97.74	95.48	96.61	95.48	96.61
PART	97.18	97.18	96.05	96.05	96.05	96.50
Average accuracy against all classifiers	96.55	97.24	96.75	96.83	96.26	96.72

($BGHJKM$) having highest classification accuracy is considered as final reduct of the dataset and used for generation of classification rules. Now wine dataset contains six conditional attributes and one decision attribute with three distinct decision classes (0, 1, 2). From the reduct, using decision matrix approach classification rules are generated. Here 60% and 40% of data are used for training and testing purpose respectively and 92% of classification accuracy is achieved. Following classification rules are generated for the wine data set, using reduct($BGHJKM$).

4. CONCLUSIONS

The proposed dimension reduction method used only the concepts of rough set theory which does not require any additional information except the decision system itself. Since, reduct generation is a NP-complete problem, so different researchers' use different heuristics to compute multiple reducts used for developing classifiers. However, using large number of reducts increases complexity of the system. Also, selecting single reduct is not always good in ideal situation for better prediction. The method tries to tradeoff between the two approaches and produces a compact set of reducts. The experimental result shows that, the accuracy given by various classifiers of the wine data set are quite high. Classification rules are also generated by considering a reduct having highest classification accuracy as single reduct for the benchmark dataset like wine and 92% of classification accuracy is achieved. Future enhancements to this work are to construction of multiple sets of classifiers from multiple reduct sets and finally ensemble them to generate an efficient classifier with better accuracy.

REFERENCES

1. M L Raymer. Dimensionality Reduction Using Genetic Algorithms, *IEEE Transactions on Evolutionary Computation*, 4(2): 164-171, 2000.
2. M A Carreira-Perpinan. A Review of Dimension Reduction Techniques, *Technical report CS-96-09, Department of Computer Science, University of Sheffield*, 1997.
3. A D Gordon. Series-Chapman and Hall/CRC

Monographs on Statistics and Applied Probability, *Classification, 2nd ed. London, U.K., ISBN: 9781584880134*, 1999.
4. S K Pal and S Mitra. Multi-Layer Perceptron, Fuzzy Sets and Classification. *IEEE Tranactions Neural Networks* 3: 683-697, 1992.
5. J Komorowski, Z Pawalk, S A Polkowski. Rough Fuzzy Hybridization: A New Trend in Decision-making, *Berlin: Springer-Verlag*, pages 3-9, 1999.
6. Z Pawlak. Rough Set Theory and Its Applications to Data Analysis, *Cybernetics and Systems*, 29: 661-688, 1998.
7. M Garey and D Johnson. A Guide to the Theory of NP-Completeness, *Computers and Intractability, Freeman, New York*, 1979.
8. R W Swiniarski. Rough Sets Methods In Feature Reduction and Classification, *In International Journal of Applied Mathematics and Computer Science*, 11(3): 565-582, 2001.
9. G Qu, S Hariri and M Yousif. A New Dependency and Correlation Analysis for Features, *IEEE Transactions on Knowledge and Data Engineering*, 17(9): 1199-1207, 2005.
10. J Novovicova. Conditional Mutual Information based Feature Selection for Classification Task, *In Proceedings of the 12th Iberoamericann Congress on Pattern Recognition*, pages 417-426, 2007.
11. A Freitas. A Genetic Programming Framework for Two Data Mining Tasks Classification and Generalized Rule Induction, *In Conference on Genetic Programming, USA*, pages 96-101, 1997.
12. D Deng and H Huang. Dynamic Reduction based on Rough Sets in Incomplete Decision Systems, *Rough Sets and Knowledge Technology, LNCS*, pages 76-83, 2007.
13. J Quinlan and R Rivest. Inferring Decision Trees Using the Minimum Description Length Principle, 80: 227-248, 1989.
14. M Hansen and B Yu. Model Selection and the Principle of Minimum Description Length, 96: 746-774, 2001.
15. J R Quinlan. The Minimum Description Length and Categorical Theories, *In Proceedings 11th International Conference on Machine Learning, New Brunswick*, pages 233-241, 1994.
16. W S Roman and H Larry. Rough Sets as a Frontend as Neural-networks Texture Classifiers, *Neuro-computing*, 36: 85-102, 2001.
17. L I Kuncheva. Combining Pattern Classifiers Methods and Algorithms, *Wiley Interscience, New York*, 2005.
18. G Fumera and F Roli. Analysis of Error-reject Trade Off In Linearly Combined Multiple Classifiers, *In Pattern Recognition*, 37: 1245-1265, 2004.
19. J Kittler and M Hatef. On Combining Classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3); 226-239, 1998.
20. A A Bakar, M N Sulaiman, M Othman and M H Selamat. Propositional Satisfiability Algorithm to Find Minimal Reducts for Data Mining, *In International Journal of Computer Mathematics*, 79(4): 379-389, 2002.
21. J Bazan, A Skowron and P Synak. Dynamic Reducts as a Tool for Extracting Laws From Decision Tables, *In Proceedings of the 8th Symposium on Me-thodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence 869, Springer-Verlag,*, pages 346-355, 1994.
22. J A Bazan. A Comparison of Dynamic and Non-Dynamic Rough Set Methods for Extracting Laws from Decision Tables. *In Rough Sets in Knowledge Discovery, Physica-Verlag, Heidelberg*, pages 321-365, 1998.
23. M J Beynon. An Investigation of Reduct Selection within the Variable Precision Rough Sets Model, *In Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing (RSCTC 2000)*, pages 114-122, 2000.
24. A T Bjorvand and J Komorowski. Practical Applications of Genetic Algorithms for Efficient Reduct Computation, *In Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied-Mathematics*, 4: 601-606, 1997.
25. A Chouchoulas and Q Shen. Rough Set-Aided Keyword Reduction for Text Categorisation, *Applied Artificial Intelligence*, 15(9): 843-873, 2001.
26. M Modrzejewski. Feature Selection Using Rough Sets Theory, *In Proceedings of the 11th International Conference on Machine Learning*, pages 213-226, 1993.
27. R Jensen, Q Shen and A Tuson. Finding Rough Set Reducts with SAT, *In Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, LNAI 3641*, pages 194-203, 2005.
28. L Polkowski. Rough Sets: Mathematical Foundations Advances in Soft Computing, *Physical*

Verlag, Heidelberg, Germany, 2002.
29. L Polkowski, T Y Lin and S Tsumoto. Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, *Studies in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg, Germany*, 56: 2000.
30. S H Nguyen. Some Efficient Algorithms for Rough Set Methods, *In Proceedings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1451-1456, 1996.
31. A Ohrn. Discernibility and Rough Sets in Medicine: Tools and Applications, *Department of Computer and Information Science. Trondheim, Norway, Norwegian University of Science and Technology*, 1999.
32. O Cordn, M J DelJesus and Herrera. Evolutionary Approaches to the Learning of Fuzzy Rule-based Classification Systems, *In Evolution of Engineering Information Systems and Their Applications, CRC Press*, pages 107-160, 1999.
33. M Dash and H Liu. Feature Selection for Classification, *Intelligent Data Analysis*, 1(3): 131-156, 1997.
34. R Jensen. Performing Feature Selection with ACO, *To Appear In Swarm Intelligence and Data Mining, Springer SCI book series*, 2006.
35. R Jensen and Q Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches, *IEEE Transactions on Knowledge and Data Engineering*, 16(12): 1457-1471, 2004.
36. A Skowron and C Rauszer. The Discernibility Matrices and Functions In Information Systems, Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, pages 331-362, 1992.
37. M Kudo and J Skalansky. Comparison of Algorithms that Select Features for Pattern Classifiers, *Pattern Recognition*, 33(1): 25-41, 2000.
38. H S Nguyen and A Skowron. Boolean Reasoning for Feature Extraction Problems, *In Proceedings of the 10th International Symposium on Methodologies for Intelligent Systems*, pages 117-126, 1997.
39. S H Nguyen and A Skowron. Searching for Relational Patterns in Data, *In Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 265-276, 1997.
40. W Siedlecki and J Sklansky. On Automatic Feature Selection, *In International Journal of Pattern Recognition and Artificial Intelligence*, 2(2): 197-220, 1988.

Dr. Asit Kr. Das is an Assistant Professor of Computer Science and Technology at Bengal Engineering and Science University, Shibpur, Howrah. He has received M.Tech Degree in Computer Science and Engineering from Calcutta University. He obtained Ph.D(Engg) Degree from Bengal Engineering and Science University, Shibpur, Howrah. His research interests include Data Mining and Pattern Recognition, Rough Set Theory, Bioinformatics *etc.*.

Ms Shampa Sengupta is an Assistant Professor of Information Technology at MCKV Institute Of Engineering, Liluah, Howrah. She has received M.Tech Degree in Information Technology from Bengal Engineering and Science University, Shibpur, Howrah. Since 2010, She has been working toward the Ph.D Degree in Data Mining at Bengal Engineering and Science University, Shibpur, Howrah.