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

Asit Kr. Das\textsuperscript{a}, Shampa Sengupta\textsuperscript{b}

\textsuperscript{a}Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howrah 711 103, West Bengal, India. Contact: akdas@cs.becs.ac.in

\textsuperscript{b}Department of Information Technology, MCKV Institute of Engineering, Liluah, Howrah 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\textsuperscript{1} \textsuperscript{2} as well as classification of data\textsuperscript{3} \textsuperscript{4}. A series of reduction algorithms\textsuperscript{5} \textsuperscript{6} were constructed for all kinds of applications based on rough set models. However, determining minimal set of attributes, called reduct, is NP-complete\textsuperscript{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\textsuperscript{8} based algorithm, dependency based\textsuperscript{9} algorithm, mutual information\textsuperscript{10} based algorithm, genetic algorithm\textsuperscript{11}, and dynamic reduction algorithm\textsuperscript{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\textsuperscript{13}–\textsuperscript{15}, the minimal reduct is preferred. However, Roman\textsuperscript{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-
(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 dataset, 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 dataset 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 construction of multiple sets of classifiers from multiple reduct sets and finally ensemble them to generate an efficient classifier with better accuracy.

REFERENCES

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


34. R Jensen. Performing Feature Selection with ACO, To Appear In Swarm Intelligence and Data Mining, Springer SCI book series, 2006.


Asit Kr. Das, Shampa Sengupta

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.