Information Sciences 275 (2014) 83–100

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A set covering based approach to find the reduct of variable precision rough set

James N.K. Liu [a,*], Yanxing Hu [a], Yulin He [b]

[a] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Machine Learning Center, Faculty of Mathematics and Computer Science, Hebei University, Baoding 071002, China

ARTICLE INFO

Article history:
Received 23 June 2011
Received in revised form 22 November 2013
Accepted 9 February 2014
Available online 18 February 2014

Keywords:
Variable precision rough set
Attribute reduct
Set covering problem
Relation matrix

ABSTRACT

Attribute reduction is one of the core problems in Rough Set (RS) theory. In the Variable Precision Rough Set (VPRS) model, attribute reduction faces two difficulties: firstly, in the VPRS model, a reduct anomaly problem may arise and it may cause an inconsistency of positive regions and decision rules after attribute reduction. Secondly, the attribution reduction problem has been proved an NP-hard problem; accordingly, we would need to find a tradeoff between calculating the minimal reduct and reducing computing complexity to avoid the combinatorial explosion problem. We propose a new approach to calculate the reduct in VPRS model. This new method focuses on calculating a β-distribution reduct while avoiding the anomaly problem in the VPRS model. The basic idea of the proposed approach is to convert the reduct problem into a Set Covering Problem (SCP) according to the positive regions in the VPRS model; and consequently, a Set-Covering Heuristic Function (SCHF) algorithm is applied to calculate the reduct after this conversion. This approach keeps the positive regions consistent after the attribute reduction and moreover, based on the SCP, the performance ratio of the proposed method to calculate the minimal reduct ranges between $\ln(|U'|) - \ln\ln(|U'|) + o(1)$ and $(1 - o(1))\ln(|U'|)$ with a computational complexity having an upper bound as $o(MN(M+N)^2)$. Finally, we demonstrate the practical application of the VPRS model using real case scenario from China's electricity power yield to verify the validity of our proposed approach. We then apply statistical evaluation to explain the economic significance of the attributes.

1. Introduction

Rough Set (RS) theory has been developed since its first introduction by Pawlak in 1982 [24–26], as one of the widely used frameworks and tools aimed to process specific information granules [4,14,27]. RS provides a formal methodology to tackle data analysis problems that involve uncertain information [25,35,39], and has brought about a widespread success in numerous artificial intelligence research fields [9,10,37]. Using concepts of lower and upper approximations in the RS model, knowledge hidden in a Decision Table (DT) can be discovered and expressed. However, in the presence of errors in a given information, or a given DT derived from a small dataset, obtained results do not perform reliably well for a larger dataset. To improve the generalizability of RS, the Variable Precision Rough Set (VPRS) model was initially proposed by Ziarko in 1993 as

[*] Corresponding author. Tel.: +852 27667273; fax: +852 27740842.
E-mail address: csnkliu@comp.polyu.edu.hk (J.N.K. Liu).

http://dx.doi.org/10.1016/j.ins.2014.02.023
0020-0255/© 2014 Elsevier Inc. All rights reserved.

one of the important extensions of classical RS model [38–40,43]. Given a threshold β, the VPRS model induces a concept of partial classification so that some errors are allowed in the classification of objects at lower approximations defined by a given table. This contrasts with the classical RS model in which the classification precision is stringent without permitting errors [9,18,42]. When applied to real complex problems containing errors and uncertain information in the data set, VPRS model comes with a superior performance [2,35,40].

In both classical RS model and VPRS model, a core problem is to find some particular subsets of the condition attributes. Attributes from such subsets are considerably redundant and they can be removed without causing deterioration of classification quality and without the induction of brief decision rules inherently found in the given tables [17,35,40]. These subsets of the condition attributes are referred to as the *reduct* of the DT. Particularly, a reduct of the DT that contains the least condition attributes is called the minimal reduct. If the removal of an attribute causes an inconsistency in classification quality degree, the attribute is deemed essential, and the full set of essential attributes of the DT is defined as the *core* of the DT.

Attribute reduct in RS model finds its effectiveness in numerous domains such as data mining, knowledge discovery, pattern recognition, and decision analysis [15,19,20,34]; subsequently, numerous studies focusing on reduct of the RS model are being conducted [29]. Finding a minimal reduct points to an NP-hard problem [32]. Many algorithms were proposed to address this problem. Currently, there are three main approaches to obtain the reduct in a RS model: (1) the attribute removing algorithm, which randomly removes attributes from the attribute collection until a reduct is obtained; (2) the heuristic attribute adding algorithm, which adds attributes to the core of a DT according to some heuristic information such as the classification quality degree [24,15,36]; (3) discernibility matrix-based algorithm introduced by Skowron in 1992 through notions of discernibility matrices and discernibility functions for describing knowledge systems [32]. The latter two algorithms are characterized by pros and cons. Though attribute adding algorithm has less computing complexity, it may result in a failure to find a reduct, and besides the calculation does not necessarily detect the minimal reduct [36]. Conversely, the algorithm based on the discernibility matrices is definitively able to find the minimal reduct with the limitation in computation complexity that requires to traversal all possible combinations of the attributes, thereby creating risks of combinatorial explosion.

Unlike in the case of classical RS model, VPRS reduct methods remain relatively unexplored mainly due to difficulties in accurately defining reduct for VPRS model. In 1992, Ziarko defined β-reduct for the VPRS model by keeping the classification quality degree consistent [43]. In 2001, Beyono analyzed the reduct anomalies of β-reduct [5] and showed that under certain circumstances, the β-reduct might lead to some incompatibility with decision rules in the reduct. Jusheng Mi, Wuzhi Wei, and Wenxiu Zhang then further proposed a new definition of the VPRS reduct, now called β-distribution reduct. With a focus on the consistency of the β-positive regions, the β-distribution reduct ensures reduct-derived decision rules compatible with those from the original DT. Reduct algorithms of the VPRS model hence warrant substantial investigations. Researchers mainly employ the revised versions of heuristic attribute adding algorithm in classical RS model using the classification quality degree as heuristic information to choose the attributes to be added into a certain subset until a reduct is found, or alternatively use the discernibility matrix to traversal through all possible combinations to find the correct combination of attributes [22]. By focusing on the consistency of the β-positive regions, the β-distribution reduct can ensure the decision rules derived from the reduct are compatible with those derived from the original DT. Reduct algorithms of the VPRS model suggest little to no investigations. Researchers mainly employ the revised versions of heuristic attribute adding algorithm in classical RS model, which uses the classification quality degree as heuristic information to choose the attributes adding into a certain subset until a reduct is found, or use the discernibility matrix to traversal through all the combinations to find a correct combination of attributes [17,21,41]. Both of these approaches inherit their shortages in RS model. However, heuristic attribute adding algorithm may fail to find a reduct; in this approach, we must first locate the core. Although the definition of the core in VPRS model varies in form [41], this may affect the final output of the heuristic attribute adding algorithm.

We herein propose an algorithm based on Set Covering Problem (SCP) to calculate the reduct in the VPRS model. The well-famous SCP with a wide application in many fields takes into account of an NP-hard problem as in the case of minimal reduct problem in the VPRS model [12,23]. SCP has a long history of investigations when compared to the length of effort in solving VPRS reduct problem [1,3,12,13,28] Specifically, some classical solutions were discussed in previous investigations [6,8,30]. We first analyze the interrelationship among positive regions, equivalent classes and decision classes to build a relation matrix for describing the knowledge of a DT in VPRS model. We then prove on the basis of relation matrix the conversion of reduct problem into a typical SCP problem. Furthermore, we show that the SCP solution is equal to a reduct of the DT in VPRS model. A Set-Covering Heuristic Function (SCHF) algorithm is finally applied to calculate the reduct after the conversion. In our proposed algorithm, we find a tradeoff between reducing computing complexity and locating the minimal reduct. In contrast to heuristic attribute adding algorithm, our method does not necessarily compute the core, leaving final result unaffected as discussed earlier; the proposed algorithm can directly focus on the positive regions to ensure the compatibility of the decision rules after the reduction. What's more, our approach circumvents the risk of failure to obtain a reduct.

In the immediate sections, we briefly revisit classical RS and VPRS models, and subsequently introduce the basic concept of SCP. In Section 3, we discuss the definition of reduct in both models. Previous algorithms are also reviewed in this section. In Section 4, we introduce our new algorithm followed by the application in real life scenario in Section 5 to reveal factors that affect China electricity power yield. Section 6 outlines our conclusion and future study avenues.