Information Sciences 254 (2014) 155–180

Contents lists available at ScienceDirect

Information Sciences

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

Quick attribute reduction in inconsistent decision tables

Min Li [a,b,c,*], Changxing Shang [a,b], Shengzhong Feng [a], Jianping Fan [a]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, PR China
[b] Graduate School of Chinese Academy of Sciences, Beijing 100080, PR China
[c] Nanchang Institute of Technology, Nanchang, Jiangxi 330099, PR China

ARTICLE INFO

Article history:
Received 28 November 2011
Received in revised form 5 May 2013
Accepted 19 August 2013
Available online 27 August 2013

Keywords:
Rough set
Attribute reduction
Inconsistent decision table
Assignment reduct
Distribution reduct
Maximum distribution reduct

ABSTRACT

This paper focuses on three types of attribute reducts in inconsistent decision tables: assignment reduct, distribution reduct, and maximum distribution reduct. It is quite inconvenient to judge these three types of reduct directly according to their definitions. This paper proposes judgment theorems for the assignment reduct, the distribution reduct and the maximum distribution reduct, which are expected to greatly simplify the judging of these three types of reducts. On this basis, we derive three new types of attribute significance measures and construct the Q-ARA (Quick Assignment Reduction Algorithm), the Q-DRA (Quick Distribution Reduction Algorithm), and the Q-MDRA (Quick Maximum Distribution Reduction Algorithm). These three algorithms correspond to the three types of reducts. We conduct a series of comparative experiments with twelve UCI (machine learning data repository, University of California at Irvine) data sets (including consistent and inconsistent decision tables) to evaluate the performance of the three reduction algorithms proposed with the relevant algorithm QuickReduct [9,34]. The experimental results show that QuickReduct possesses weak robustness because it cannot find the reduct even for consistent data sets, whereas our proposed three algorithms show strong robustness because they can find the reduct for each data set. In addition, we compare the Q-DRA (Quick Distribution Reduction Algorithm) with the CEBARKNC (conditional entropy-based algorithm for reduction of knowledge without a computing core) [43] because both find the distribution reduct by using a heuristic search. The experimental results demonstrate that Q-DRA runs faster than CEBARKNC does because the distribution function of Q-DRA has a lower calculation cost. Instructive conclusions for these reduction algorithms are drawn from the perspective of classification performance for the C4.5 and RBF-SVM classifiers. Last, we make a comparison between discernibility matrix-based methods and our algorithms. The experimental results indicate that our algorithms are efficient and feasible.

1. Introduction

Rough set theory is a powerful mathematical tool introduced by Pawlak [23] to address imprecise, incomplete or vague information. Many researchers have contributed to its development and applications [4,7,9,10,16,17,24–27,33–35,37,39–41,45]. One fundamental aspect of rough set theory is attribute reduction in information systems (ISs), which is selecting or reserving those attributes that provide the same information for classification purposes as the entire set of available attributes. There are many types of attribute reductions in the area of rough sets [1,5,6,8,13,15–22,33–35,38,42–44,46–54]. Pawlak proposes the classic attribute reduction, which is intended to preserve the deterministic information with respect to

* Corresponding author. Tel.: +86 13607002079.
E-mail address: liminghuadi@hotmail.com (M. Li).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.08.038

decision attributes of a decision table and is therefore often applied in extracting deterministic decision rules from the decision tables. However, most of the decision tables are not consistent because of various factors such as noise in the data, lack of critical knowledge, compact representation, and prediction capability. There is no guarantee for such classic attribute reduction to preserve non-deterministic decision information in an inconsistent decision table. Regarding this concern, Kryszkiewicz presents two types of attribute reduction for inconsistent decision tables: assignment reduction and distribution reduction [11,12]. Assignment reduction maintains unchanged the possible decisions for an arbitrary object in an inconsistent decision table. In comparison, distribution reduction is a more complete knowledge reduction and is characterized by preserving the class membership distribution for all of the objects in an inconsistent decision table. In other words, the distribution reduction preserves not only all of the deterministic information but also the non-deterministic information of an inconsistent decision table. Yao thinks that the partition based on the membership distribution vector is finer and more complex, which allows the distribution reduction to preserve the quality of the decisions [46]. However, it could be a concern that the distribution reduction has strict requirements, and the decision rules derived from distribution reduction are usually less compact and more complicated. For this reason, Zhang et al. have proposed the maximum distribution reduction [51,52]. It maintains unchanged the maximum decision classes for all of the objects in a decision table, which is seen as a good compromise between the capability of preserving information with respect to decisions and the compactness of the derived rules. In Ref [47], Ye et al. have presented another extended type of attribute reduction, called M-reduct. M-reduct has stricter requirements than the maximum distribution reduction in that M-reduct rigidly retains the membership degree to the maximum decision class for each object of the decision table. This characteristic makes M-reduct more susceptible to noise contaminated data sets than the maximum distribution reduction because a small amount of noisy data can cause the M-reduct to include more attributes.

To find the three types of reducts (the assignment reduct, the distribution reduct and the maximum distribution reduct), Zhang et al. have utilized the discernibility matrixes with respect to those reducts and have obtained the corresponding Boolean function, called the discernibility function [36,51]. This function is reduced by using the distribution and the absorption laws. And then all of the reducts are generated by finding all of the prime implicants of the function. The attribute reduction based on the discernibility matrix has been extensively researched [2,13,19,21,28]. For example, a type of attribute reduction called a lower approximation reduct and an upper approximation reduct is presented [28]. This type of attribute reduction preserves the lower/upper approximation distribution of a target decision. In fact, the lower approximation reduct is equivalent to the relative positive region reduct, and the upper approximation reduct is equivalent to the assignment reduct in inconsistent (complete) decision tables. The discernibility matrixes associated with the two approximation reductions are examined as well. In paper [2], Chen et al. have constructed tolerance granules by defining a cover of the universe and have proposed a discernibility matrix-based reduction algorithm for computing the relative reducts of consistent and inconsistent covering decision systems. Similarly, the reduction approach based on discernibility matrixes is also found in [21], where the region preservation reduct (equivalent to the position region reduct), decision preservation reduct (equivalent to the assignment reduct) and relationship preservation reduct are discussed in detail. Then, three distinct definitions of discernibility matrixes are defined to find these types of reducts. More recently, a relatively systematic study of attribution reduction in inconsistent incomplete decision tables is presented in [19], where five types of discernibility function-based approaches are proposed to identify a specific type of reduct. Although discernibility matrix-based methods can find all of the reducts, the conversion from conjunction normal form to disjunction normal form constitutes an NP-hard problem. When the data set has many attributes, these discernibility matrix-based methods become not feasible because the matrix contains too many candidates. Therefore, heuristic methods are desirable. In [18], Meng and Shi developed a fast attribute reduction algorithm for incomplete decision systems based on tolerance relation rough sets. The complexity of this algorithm is no more than $O(|C|^2|U|\log|U|)$, which means that the algorithm can be used for attribute reduction in large-scale decision systems. However, the proposed approach is suitable only for computing the positive region-based reducts of an inconsistent incomplete decision system.

It is quite inconvenient to judge the three types of reducts, namely the assignment, the distribution and the maximum distribution reducts, directly according to their definitions because their definitions are rather complex. In this paper, we propose judgment theorems for the assignment reduct, the distribution reduct and the maximum distribution reduct, which are expected to greatly simplify the judging of these three types of reducts. On this basis, we derive three novel types of attribute significance measures and construct the Quick Assignment Reduction Algorithm, the Quick Distribution Reduction Algorithm, and the Quick Maximum Distribution Reduction Algorithm, so as to correspond to these three types of reducts. These three algorithms have similarities to the well-known reduction algorithm of QuickReduct [9,34] in a forward greedy search form, with respect to time and space complexities. However, QuickReduct only finds the positive region reducts, which are often applied in consistent decision tables. A series of comparative experiments with twelve UCI data sets (including consistent and inconsistent decision tables) show that the QuickReduct possesses weak robustness because it cannot find the reduct, even for consistent data sets, whereas our proposed three algorithms show strong robustness and they find the reduct for each data set. At the same time, some instructive conclusions for these reduction algorithms are drawn from the perspective of classification performance on C4.5 and RBF-SVM classifiers. The CEBARKNC [43] proves to be a reduction algorithm with efficiency because it is applied to find the distribution reduct for inconsistent decision tables. Additionally in our experiments, we compare the Quick Distribution Reduction Algorithm with the CEBARKNC because both find distribution reducts. The results show that both have a high degree of similarity: they find 9 matching reducts in the 12 data sets. Nevertheless, the experimental results prove that the simple calculation of our distribution function enables the