International Journal of Approximate Reasoning 82 (2017) 56–80

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Quick general reduction algorithms for inconsistent decision tables

Ge Hao [a,b], Li Longshu [b,c,*], Xu Yi [b,c], Yang Chuanjian [d]

[a] School of Electronic and Electrical Engineering, Chuzhou University, Chuzhou 239000, People's Republic of China
[b] Key Laboratory of Computation Intelligence and Signal Processing of Education Ministry, Anhui University, Hefei 230601, People's Republic of China
[c] School of Computer Science and Technology, Anhui University, Hefei 230601, People's Republic of China
[d] School of Computer and Information Engineering, Chuzhou University, Chuzhou 239000, People's Republic of China

ARTICLE INFO

Article history:
Received 2 August 2015
Received in revised form 25 November 2016
Accepted 26 November 2016
Available online 2 December 2016

Keywords:
Rough set
Attribute reduction
Discernibility matrix
Relative discernibility

ABSTRACT

In the Pawlak rough set model, attribute reduction plays one of the important roles, and the preservation of different properties of the original decision table leads to different types of reduct definitions, such as relative relation reduct, positive-region reduct, distribution reduct, maximum distribution reduct, and assignment reduct. However, there are no general quick reduction approaches for obtaining various types of reducts; this motivated us to conduct the present study. We first establish a unified decision table model for five representative reducts in inconsistent decision tables, study the relative discernibility and relative discernibility reduct for the general decision table, and derive the corresponding properties. Then, two general reduction algorithms (GARA-FS$_\triangle$ and GARA-BS$_\triangle$) from the viewpoint of the relative discernibility in inconsistent decision tables are presented. Subsequently, to increase the efficiency of algorithms, two quick general reduction algorithms (QGARA-FS$_\triangle$ and QGARA-BS$_\triangle$) are proposed mainly by reducing the sort times to increase the efficiency of reduction algorithms. Finally, a series of experiments with UCI data sets are conducted to evaluate the effectiveness and performance of the proposed reduction algorithms.

1. Introduction

Rough set theory was introduced by Polish mathematician Pawlak [24,25] in 1982, which is a valid mathematical theory to handle imprecise, uncertain, and vague information. It has been widely applied in many fields such as machine learning [33], data mining [15], intelligent data analyzing [4], and control algorithm acquisition [34]. Some of the features in huge-volume and high-dimensional data sets are irrelevant or redundant, which typically deteriorates the performance of machine-learning algorithms [29]. Attribute reduction is one of the key topics in the rough set theory, which can find a subset of attributes that provides the same description, discernibility, or classification ability as the original conditional attribute set. In general, attribute reduction uses a preprocessing procedure to reduce the complexity of data mining or knowledge discovery. In recent years, attribute reduction has drawn wide attention, and many researchers have focused on two aspects: various types of reduct definitions based on different criteria and quick attribute reduction algorithms.

* Corresponding author.
E-mail addresses: togehao@126.com (H. Ge), Lilongshu_ahu@126.com (L. Li), xuyi1023@126.com (Y. Xu), tocjy474@126.com (C. Yang).

http://dx.doi.org/10.1016/j.ijar.2016.11.016
0888-613X/

The different purposes for preserving a certain property embedded in original decision table will result in different notions and results of reduct. The positive-region reduct proposed by Pawlak is the classical concept of reducts [24,25], which focuses on preserving the positive region unchanged. Kryszkiewicz presented five notions of reducts for inconsistent decision tables. In fact, there only existed two types of reducts: the assignment reduct and the distribution reduct [8,9]. The assignment reduct maintains unchanged property of the possible decisions for an arbitrary object in an inconsistent decision table. In comparison, the distribution reduct is a more complete knowledge reduct that is characterized by preserving the class membership distribution for all the objects in an inconsistent decision table. In other words, the distribution reduction preserves not only all the deterministic information, but also the nondeterministic information of an inconsistent decision table. On the basis of Kryszkiewicz's work, Zhang et al. proposed the maximum distribution reduct [41,42]. It maintains unchanged property of the maximum decision classes for all 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 refs. [41] and [42], Zhang et al. provided discernibility matrix-based reduction methods for distribution reduct, assignment reduct, and maximum distribution reduct. Ye et al. [40] presented M-reduct, which has stricter requirements than the maximum distribution reduct, in which 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. Miao et al. [20] presented three types of reducts for inconsistent decision tables: the region preservation reduct, the decision preservation reduct, and the relationship preservation reduct, and common definitions of relative reduct and discernibility matrices for three reducts were discussed. Liu et al. [17] presented a method to convert these types of reducts (the distribution reduct, the maximum distribution reduct, and the generalized decision reduct) into traditional reducts, and designed an efficient algorithm for the traditional reduct. Zhou et al. [45] proposed 13 typical forms of objective functions for attribute reduction in the complete decision table. However, among the 13 types of reducts for the decision table, it has been proved that there are only six and two intrinsically different reducts for inconsistent and consistent decision tables, respectively. Meng et al. [23] have recently presented a relative systematic study of attribute reduction in inconsistent incomplete decision tables, where five types of discernibility function-based approaches are proposed to identify a specific type of reducts.

The other research of attribute reduction with rough set mainly concentrates on developing feature selection algorithms to find reducts according to different models of rough sets [1–3,10,22,32,34,35], and these methods normally can be grouped into two classes: discernibility matrix-based reduction method and heuristic reduction method. Although discernibility matrix-based method can find all of the reducts, the conversion from conjunction normal form to disjunction normal form constitutes a nondeterministic polynomial (NP)-hardness problem. When data sets have many attributes and objects, discernibility matrix-based method will become nonfeasible, as the matrix contains too many candidates. The heuristic reduction method is one of the desirable methods for overcoming the drawbacks of the discernibility matrix approaches. Shen et al. [28–30] presented QuickReduct algorithm, which started with empty subset, and then added the most significant feature into the candidate set of reduct until dependency reached the maximum value in the data set. Miao et al. [21] introduced an information representation to the rough set theory and proposed a heuristic algorithm based on mutual information for attribute reduction. Liu et al. [16] proposed a new conditional information entropy. By using this entropy, the new significance of an attribute was defined, and an efficient heuristic algorithm for computing knowledge reduct was designed. Yang et al. [38] proposed a heuristic feature selection algorithm in incomplete decision tables, which keeps the positive region of target decision unchanged. Liang et al. defined a new information entropy to measure the uncertainty of the incomplete information systems [14] and applied the corresponding conditional entropy to reduce redundant features [13]. Meng et al. [22] developed a heuristic attribute reduction algorithm based on tolerance relation rough set, which was beneficial to attribute reduction in large-scale decision systems.

To further increase the efficiency of reduction algorithms, many researchers studied acceleration mechanisms. Because partition (or equivalence) relation is one of the important and primitive notions in the reduction approaches, the computation of equivalence class is one of the key steps for attribute reduction. Hence, Liu et al. [18] proposed the approach of computing the equivalence classes by quick sort, and designed a reduction algorithm based on the positive region with time complexity $O(|C|^2|U|\log|U|)$, where C represents the conditional attribute set, U denotes the object set, and $|\cdot|$ denotes the cardinality of set. Xu et al. [36] computed equivalence classes by radix sort, which improved the performance of attribute reduction algorithms and made heuristic reduction algorithm's time complexity to reduce $\max(O(|C||U|), O(|C|^2|U/C|))$. Qian et al. [26] also presented a counting sort algorithm to deal with inconsistent decision tables and compute positive regions and core attributes, and a hybrid attribute reduction algorithm based on the relationship between indiscernibility and discernibility was proposed with time complexity no more than $\max(O(|C||U|), O(|C|^2|U/C|))$. To further reduce computational time, by using four types of common heuristic reduction algorithms, Qian et al. [27,28] studied an accelerator strategy for the positive-region reduct and three types of entropy reducts in the complete and incomplete decision tables. The heuristic reduction methods based on the accelerator could significantly decrease the time consumed and obtain the same reduct as their original methods. On the basis of the above research, Liang et al. [12] developed a new accelerator that simultaneously decreased the size of universe and reduced the number of attributes in each iteration process of attribute reduction to further accelerate attribute reduction efficiency. In addition, Li et al. [11] proposed judgment theorems for the assignment reduct, distribution reduct, and maximum distribution reduct, derived three new types of attribute significance measure based on judgment theorems, and constructed three quick algorithms corresponding to the three types of reduct.