Information Sciences 330 (2016) 226–244

Contents lists available at ScienceDirect

Information Sciences

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

On quick attribute reduction in decision-theoretic rough set models

Zuqiang Meng[a,*], Zhongzhi Shi[b]

[a] College of Computer, Electronics and Information, Guangxi University, Nanning 530004, PR China
[b] The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 6 December 2014
Revised 23 September 2015
Accepted 27 September 2015
Available online 17 October 2015

Keywords:
Decision-theoretic rough set model
Attribute reduction
Efficiency
Rough set theory
Equivalence class
Super-reduct

ABSTRACT

Compared with the Pawlak rough set model, decision-theoretic rough set (DTRS) models have relatively powerful data processing capability for complex and low quality data. However, the non-monotonicity of the criteria used in DTRS models causes many problems, particularly the inefficiency of attribute reduction algorithms, which greatly narrows their applications. Few systematic studies have been reported on how to construct efficient reduction methods for DTRS models. This paper discusses this problem and constructs an efficient reduction algorithm for DTRS models incorporating three aspects: taking advantage of common characteristics of data sets and employing division techniques, a fast division approach to computing equivalence classes is proposed, which is the most frequently used and the most time-consuming basic operation in the reduction algorithm; extracting a "monotonic ingredient" from decision systems, an effective heuristic function is constructed for the reduction algorithm, which can guide searches for the algorithm and has better ability to find shorter super-reducts; a novel algorithm is proposed to search the power set space of a given super-reduct, which has a shorter average search length than the existing backtracking methods. Finally, the reduction algorithm is proposed that combines these improvements. Experimental results demonstrate that these improvements are effective and the proposed reduction algorithm is relatively efficient.

1. Introduction

Attribute reduction is one of the most important topics in rough set theory, which is a useful mathematical tool for data analysis and has been widely used in many fields [2,7,16,23,27,34,36,40,41,44]. In rough set theory, the definition and interpretation of attribute reduction are mainly dependent on construction methods of the lower and upper approximations. In the Pawlak rough set model [28], an equivalence class belongs to the lower approximation of a subset if and only if all objects in the equivalence class belong to the subset. However, this requirement is too strict for some cases. If very few objects in an equivalence class do not belong to a given subset because of some subjective or objective factors, such as noisy data, the equivalence class cannot be assigned to the lower approximation of the subset. Therefore, this approach is not suitable for noisy data, which is inevitable in many applications. To solve this problem, many scholars incorporate probabilistic approaches into the Pawlak model by introducing thresholds, and have developed many extended models, such as variable precision rough set

* Corresponding author. Tel./fax:+86 771 3232214.
E-mail address: zqmeng@126.com (Z. Meng).

http://dx.doi.org/10.1016/j.ins.2015.09.057
0020-0255/

models [59], parameterized rough set models [16,30,37], probabilistic rough set models [4,48,50,57,58], and Bayesian rough set models [3,38]. Yao et al. conducted a great deal of research work in this aspect and developed the well-known decision-theoretic rough set (DTRS) model [47–49,51–53]. The DTRS model has been successfully applied to many areas [10,14,15,18,20,39,46,55] and is now increasingly attracting more and more attention [6,9,15].

Probabilistic models, including variable precision rough set models, parameterized rough set models, etc., generally lack physical interpretation of the required thresholds and systematic methods for setting the thresholds [49]. In these models, setting the thresholds relies on subjective experience or tedious trial-and-error approaches. Yao et al. introduced the Bayesian decision procedure into probabilistic models to resolve the above problems, and subsequently developed the DTRS model, which provides not only theoretical support for the setting of the thresholds but also physical interpretation for the required thresholds. Furthermore, different probabilistic models can be derived from the DTRS model [53].

Therefore, the DTRS model is an important typical probabilistic rough set model, and has been widely studied. Zhao et al. [56] analyzed the definitions of attribute reduction in DTRS models with emphasis on the probabilistic regions. They considered two groups of studies on attribute reduction, region preservation reduction and decision preservation reduction, and showed that the former is an optimization problem while the latter is a decision problem. They also provided corresponding definitions of attribute reduction in DTRS models. Yao et al. [53] analyzed the properties of different criteria, such as decision monotonicity, confidence, coverage, generality, and cost, and proposed a general definition of a probabilistic attribute reduct that allows the definition of a proper attribute reduct for the particular application. Li et al. [11] analyzed the monotonicity of the positive region in the Pawlak rough set and DTRS models, and then proposed a positive region expanding reduct based on decision monotonicity. However, decision monotonicity is a subjective criterion. Jia et al. [5] proposed a definition of attribute reduction, minimum cost attribute reduct, for DTRS models, which is formulated as an optimization problem with the objective of minimizing the decision cost. The proposed reduction provided a simple way to evaluate and interpret region preservation attribute reduction in DTRS models. Li et al. [12] proposed a non-monotonic attribute reduction for the DTRS model, which allowed the positive region to expand rather than remain constant. It can be found that, in DTRS models, different types of reducts can preserve different properties, and each property is evaluated by a specific criterion or measure. These criteria are usually divided into two categories: qualitative criteria, such as positive region, and quantitative criteria, such as quality of classification and cardinality of the positive region. Each category gives rise to a different definition for attribute reduction, and correspondingly there are two types of reduction definitions: qualitative definition and quantitative definition [5,22], which are united in the general definition of probabilistic attribute reduction proposed by Yao et al. [53].

In DTRS models, probabilistic methods are used to construct lower and upper approximations, which allow a certain degree of error rate, hence both qualitative criteria and quantitative criteria are not necessarily monotonic with respect to set inclusion of attributes. In fact, monotonic region-based criteria, which are widely used in DTRS models, are unavailable to date [6,22]. Thus, the definitions of attribute reduction are generally not equivalent in DTRS models. These factors make the problem of attribute reduction more complex. Although there are many successful efficient reduction methods [19,32,43] for the Pawlak model, they cannot be directly applied to DTRS models, and it remains difficult to construct an efficient method to find a minimal reduct in DTRS models.

The objective of attribute reduction is to reduce the number of redundant attributes and preserve desirable related properties. In DTRS models, a reduct is a minimal subset of attributes that satisfies the jointly sufficient and individually necessary conditions [53]. An attribute subset that satisfies the jointly sufficient condition is a super-reduct [26,45,56]. Because of the non-monotonicity of the criteria used in DTRS models, all non-empty subsets of a super-reduct must be checked to guarantee that the final reduct satisfies the individually necessary condition.

Suppose B is a super-reduct, then the number of all non-empty subsets of B is $2^{|B|}-1$, which is an exponential function with respect to $|B|$. To solve this problem, an alternative method of defining reducts is to relax the individually necessary condition as follows [5,56]: check all subsets $B-\{a\}$ for all $a \in B$ rather than all subsets $B' \subseteq B$, where the number of the former is $|B|$ while that of the latter is $2^{|B|}-1$. Obviously, if this definition is adopted, the efficiency of reduction algorithms can be greatly improved. Unfortunately, there are possibly redundant attributes in such a reduct and corresponding reduction algorithms are incomplete. Therefore, strictly speaking, a subset that satisfies the jointly sufficient condition and the relaxed individually necessary condition is not a reduct but an approximate reduct [5], which is still a super-reduct. Only a subset that satisfies both the jointly sufficient condition and the individually necessary condition is actually a reduct, and designing efficient algorithms for finding such a reduct in DTRS models is the main task of this paper. We also notice that many studies [1,17,21,35] have been conducted on constructing fast reduction methods in the field of rough set in recent years, but few of them are directly oriented to DTRS models. For example, paper [21] proposed two fast incremental algorithms for computing rough approximations in set-valued decision systems; paper [17] used bucket techniques to construct a quick attribute reduct algorithm for neighborhood rough set model; paper [35] proposed a fast approach to attribute reduction in incomplete decision systems by designing new attribute measures to evaluate the quality of candidate attributes. In our view, there are several key problems to be addressed in designing efficient reduction methods for DTRS models.

Firstly, a serious problem of attribute reduction in DTRS models is the inefficiency of reduction algorithms, because reduction algorithms need to check all non-empty subsets of a super-reduct, which leads to exponential complexity in the worst case. Scholars generally use the backtracking technique to improve the efficiency of finding a reduct in a given super-reduct [22,26,45]. However, our study shows that these backtracking methods can be further improved, and we propose a novel and more efficient search method to find a minimal reduct in a given super-reduct. This is the first contribution of this paper.