International Journal of Approximate Reasoning 55 (2014) 867–884

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Updating attribute reduction in incomplete decision systems with the variation of attribute set

Wenhao Shu [a], Hong Shen [a,b,*]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China
[b] School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form 21 September 2013
Accepted 24 September 2013
Available online 9 October 2013

Keywords:
Rough set
Attribute reduction
Incomplete decision systems
Positive region
Dynamic attribute set

ABSTRACT

In rough set theory, attribute reduction is a challenging problem in the applications in which data with numbers of attributes available. Moreover, due to dynamic characteristics of data collection in decision systems, attribute reduction will change dynamically as attribute set in decision systems varies over time. How to carry out updating attribute reduction by utilizing previous information is an important task that can help to improve the efficiency of knowledge discovery. In view of that attribute reduction algorithms in incomplete decision systems with the variation of attribute set have not yet been discussed so far. This paper focuses on positive region-based attribute reduction algorithm to solve the attribute reduction problem efficiently in the incomplete decision systems with dynamically varying attribute set. We first introduce an incremental manner to calculate the new positive region and tolerance classes. Consequently, based on the calculated positive region and tolerance classes, the corresponding attribute reduction algorithms on how to compute new attribute reduct are put forward respectively when an attribute set is added into and deleted from the incomplete decision systems. Finally, numerical experiments conducted on different data sets from UCI validate the effectiveness and efficiency of the proposed algorithms in incomplete decision systems with the variation of attribute set.

1. Introduction

Reducing dimensionality is often carried out as data preprocessing to improve recognition accuracy in pattern recognition, data mining, machine learning [7,17–19]. This process is called as feature selection. Feature selection based on rough set theory is also called attribute reduction. As an important data preprocessing step, the main goal of attribute reduction is to find a minimal subset of attributes, which can unambiguously classify objects of the same domain as the full original set of available attributes [2,31,40]. The classical rough set-based attribute reduction presents a systemic theoretic framework of feature selection for complete data. However, since data are generally imprecise, tend to be noisy, it is most often the case that the values of attributes may be missing or have multiple values on some other attributes in real-world information systems [5,9,33], i.e., there exist objects whose attribute values are unknown, and this is where traditional rough set theory encounters a problem. It is not possible in the traditional theory to say whether two attribute values are similar. For example, two close values may only differ as a result of uncertainty and incompleteness of information, but in the traditional rough set theory, they are considered to be as different. A possible way of combating this would be to propose extensions to the traditional theory. Different generalizations have been recently considered [11,38,42], they are based on similarity or

* Corresponding author at: School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China.
E-mail addresses: 11112084@bjtu.edu.cn (W. Shu), hshen@bjtu.edu.cn (H. Shen).

0888-613X/$ – see front matter
http://dx.doi.org/10.1016/j.ijar.2013.09.015

tolerance relations in the literatures [1,8,10,15,16]. An early extension of rough sets that can directly deal with incomplete data presented by Kryszkiewicz in [8] is under a tolerance relation. Similar to the way in [8], a missing attribute value is interpreted as any known value of a corresponding attribute in this paper.

Since attribute reduction is as one of the many successful applications in rough set theory, but due to the presence of missing attribute values, the attribute reduction in incomplete data sets becomes more complex. A key reason is that the classes of all elementary set are no longer a partition but a coverage of the universe. Generally speaking, attribute reduction from incomplete data usually starts from incomplete information systems. If condition attributes and decision attributes in an incomplete information system are distinguished, it is called an incomplete decision system. Given an incomplete decision system, an attribute reduct is a subset *Red* of the attributes which have the same information content as the full condition attribute set *C* [32]. There are lots of researches related to attribute reduction from incomplete decision systems. Many heuristic attribute reduction methods are developed, and these methods can be broadly classified into three categories: One based on positive region [12,16,36,37], the other based on discernibility matrix [14,15], and another based on entropy [10,13]. And in order to accelerate a heuristic process of attribute reduction, a theoretic framework named positive approximation is proposed [3]. Based on the proposed accelerator, the efficiency of some general attribute reduction algorithms is improved. By using the positive region-based attribute reduction algorithms, knowledge hidden in data can be represented as certain rules. The certainty embodied in the positive region is associated with greater importance in scientific analysis, almost all positive region-based algorithms adopt an approach to minimization that employs the information contained within the lower approximation of a set. In this regard, the positive region-based attribute reduction algorithm is employed in this paper. The objective is to provide a methodology for attribute reduction in incomplete decision systems as well as to further enrich the theory and method of reduction in incomplete decision systems and provide support for extending related reduction methods.

However, due to dynamic characteristics of data collection in decision systems, correspondingly, the set of attribute reduct generated from decision systems is sensitive to changes in the systems [4], which needs to update for dynamic data mining and related tasks. Since a decision system consists of attribute set, object set and the domain of attributes values, researches related to attribute reduction from dynamic decision systems can be broadly classified into three classes. The first class is the study on updating attribute reduction caused by the variation of attribute set, the second class is the study on updating attribute reduction caused by the variation of object set, and the third class is the study updating attribute reduction caused by the variation of attribute values. In real-life applications, dynamic changing of attribute set often happens in most situations. This paper concentrates on attribute reduction for dynamic variation of an attribute set in incomplete decision systems. Generally speaking, a dynamic change of attribute set in the incomplete decision system includes two aspects, i.e., adding attribute set and deleting attribute set. Here are some specific explanations for the two aspects. One situation is that one initially gather many attributes that are potentially useful. However, some gathered attributes may not be useful in later task. Deleting the attributes that have no contribution to classification or decision will be beneficial to the whole data mining process. For example, there is an incomplete decision system about the disease flu, which is described by numbers of attributes (headache, muscle pain, temperature, age, weight, blood pressure and family history), however, there are some attributes with many missing attribute values, and some attributes do not meet doctor's need. To improve a better performance, it often deletes those attributes from the original attribute set that may not be useful or relevant to doctor's decision making. The other situation is that collecting attribute set may increase gradually according to the need of real-life application, though it usually is not necessary to require originally. And adding some necessary attributes into the original attribute set will improve the accuracy of classification in an effective way. For example, for the incomplete decision system about the disease flu, as new examination tools growing in some attributes with missing attribute values, more attributes are added to the original attribute set to capture new necessary information on disease flu.

Whenever adding or deleting attributes, one may retrain the incomplete decision system from scratch to obtain a new reduct, which is known as a non-incremental manner. However, the non-incremental manner becomes very costly or even infeasible to run repeatedly as the number of attributes grows. To overcome this deficiency, one can apply an incremental analytic approach. The essence of incremental analytic approach is to allow the learning process to employ an on-going and progressive manner instead of a one-off way. The researches on updating knowledge incrementally have shown its importance in many areas, such as clinical decision making, intrusion detection, stock evaluation, and text categorization. Some incremental learning methods with respect to rough set theory have been proposed [6,21–30]. However, most of incremental learning approaches in rough set theory focus on dynamic complete decision systems instead of dynamic incomplete decision systems. Therefore, updating attribute reduction in incomplete decision systems with the variation of attribute set is discussed in this paper.

The remainder of this paper is structured as follows: Section 2 summarizes some related work on attribute reduction in incomplete decision systems and some incremental updating approaches in rough set theory. Section 3 reviews the theoretical background of rough set theory and relevant definitions involved in this paper. Section 4 establishes an incremental manner to calculate new positive regions with the variation of attribute set in incomplete decision systems. In Section 5, by incorporating the calculated positive region into the computation of attribute reduction, the corresponding attribute reduction algorithms are designed respectively when an attribute set is added into and deleted from an incomplete decision system. In Section 6, extensive experimental studies are conducted, which demonstrate the effectiveness and efficiency of the proposed algorithms. Finally, Section 7 concludes the paper and discusses the prospects for further work.