Data & Knowledge Engineering xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

An incremental approach to attribute reduction from dynamic incomplete decision systems in rough set theory

Wenhao Shu [a], Wenbin Qian [b,c,*]

[a] School of Information Engineering, East China Jiaotong University, Nanchang 330013, PR China
[b] School of Software, Jiangxi Agriculture University, Nanchang 330045, PR China
[c] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 3 August 2013
Received in revised form 15 May 2015
Accepted 22 June 2015
Available online xxxx

Keywords:
Attribute reduction
Positive region
Dynamic incomplete decision systems
Knowledge acquisition
Rough sets

ABSTRACT

Attribute reduction is an important preprocessing step in data mining and knowledge discovery. The effective computation of an attribute reduct has a direct bearing on the efficiency of knowledge acquisition and various related tasks. In real-world applications, some attribute values for an object may be incomplete and an object set may vary dynamically in the knowledge representation systems, also called decision systems in rough set theory. There are relatively few studies on attribute reduction in such systems. This paper mainly focuses on this issue. For the immigration and emigration of a single object in the incomplete decision system, an incremental attribute reduction algorithm is developed to compute a new attribute reduct, rather than to obtain the dynamic system as a new one that has to be computed from scratch. In particular, for the immigration and emigration of multiple objects in the system, another incremental reduction algorithm guarantees that a new attribute reduct can be computed on the fly, which avoids some re-computations. Compared with other attribute reduction algorithms, the proposed algorithms can effectively reduce the time required for reduct computations without losing the classification performance. Experiments on different real-life data sets are conducted to test and demonstrate the efficiency and effectiveness of the proposed algorithms.

1. Introduction

Rough set theory [6,48–50], proposed by Pawlak, is an extension of set theory. It is a powerful mathematical tool for knowledge discovery, decision analysis, data mining, machine learning, and so on [9,37,41,43]. For some data mining and machine learning tasks, a large number of features are stored in data sets in various practical applications [1,2,13]. It has been observed that an excessive number of features may cause deterioration of the results when using data mining tools for knowledge discovery, because redundant and irrelevant features are highly confusing in the knowledge learning process. Feature selection becomes an essential task before decision-making analysis. To acquire more compact decision rules, some feature subsets in the body (the condition part) of the rules are needed. On feature selection, a special theoretical framework to select useful features is Pawlak's rough set theory [3,14,44]. The main advantage of rough set theory is that it requires no preliminary or additional information about data. It works by making use of the data only. This is a major difference compared to other methods that require supplementary knowledge such as probabilistic distribution in statistical methods [38], grade of membership in fuzzy set theory [12,42] and basic probability assignment in Dempster–Shafer theory [18]. Feature selection based on rough set theory is also called attribute reduction. *Attribute reduction* is a process to find the optimal subset of attributes that retains the same discriminatory power of the whole attribute set, to eliminate the attributes that are unimportant or irrelevant to the target concept. It serves as a pre-processing stage to effectively eliminate redundant attributes without affecting the classification

* Corresponding author at: Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, PR China.
E-mail addresses: 11112084@bjtu.edu.cn (W. Shu), qianwenbin1027@126.com (W. Qian).

http://dx.doi.org/10.1016/j.datak.2015.06.009
0169-023X/

performance. Most existing attribute reduction algorithms rely on the information gathered from the positive region [17,22,25,53]. A positive region contains all the objects that can be classified into the decision attribute set using the information in the condition attribute set. The *positive region* is defined as the union of the lower approximations. The lower approximation is the set of the domain objects that certainly belongs to the concept of interest. By using the positive region based attribute reduction algorithms, certain information can be discovered and certain rules can be derived. The 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.

Attribute reduction based on rough set theory starts from an information system that contains data about the objects of interest, which are characterized by a finite set of attributes. For an information system, if condition attributes and decision attributes are distinguished from each other, it is called a decision system. In fact, there may be multiple reducts for a given decision system. However, it is enough to find one reduct in most applications. It has been proven that finding the minimal reduct is NP-hard [10]. Thus, various heuristic approaches to attribute reduction have been suggested by many authors. From the perspective of different attribute criteria in rough set theory, the heuristic attribute reduction algorithms encountered in the literature can be mainly categorized into three representative types depending upon the measure of attribute selection utilized in their design: positive region-based attribute reduction [17,25,53], discernibility matrix-based attribute reduction [7,10,23], and entropy-based attribute reduction [9,19,20]. The main differences of these algorithms lie in the metrics that are used to evaluate the quality of candidate attributes to find optimal solutions. Work on attribute reduction in classical rough set theory model has focused on complete decision systems, i.e., the values of attributes are complete. However, in many real-world tasks, it may occur that some of the attribute values for an object are incomplete (missing) due to the restriction of access, the errors of measurement and so on. However, rules must be extracted from incomplete data, which motivates many researchers to study various approaches to address incomplete decision systems [8,11,16,17,19–24,30,39,40]. Generally speaking, according to whether a given decision system has missing attribute values, it can be classified into two categories: complete decision systems and incomplete decision systems. Two main semantics for incomplete attribute values are systematically studied in Refs. [16,52]: the absent value semantics and the missing value semantics. With the absent value semantics, the incomplete attribute values are not accessible, although they were known originally, for a variety of reasons, e.g., they were mistakenly erased or forgotten to be entered into the data set. In the missing value semantics, the incomplete attribute values are irrelevant or unimportant. They are replaced by all of the values from the domain of the attribute, which can be classified in spite of the fact that some attribute values are not known. In this context, such incomplete attribute values can be further divided into three categories by different interpretations: "do not care" conditions, restricted "do not care" conditions and attribute-concept values. The way in which incomplete attribute values are considered as missing value semantics is relatively representative [8,11,17,19–24,30], and it can be more easily modified to solve the absent value semantics. Therefore, we interpret an incomplete attribute value as any possible value of each attribute according to the missing value semantics in incomplete decision systems. Therefore, the equivalence relation that is suitable for complete data in classical rough set theory is extended to a tolerance relation for incomplete data.

However, real-world decision systems such as clinical decision making systems, intrusion detection systems, stock evaluation systems, and some real-time application systems, often vary dynamically over time. It is not surprising that the non-incremental approach in attribute reduction may not be applied to such systems. As it usually needs to retrain the dynamic systems as new ones, large amounts of computational time and memory space are needed for recomputations. Especially for large-scale dynamic decision systems, the non-incremental approach becomes very costly or even intractable. Therefore, it is desirable to develop new analytic techniques to address such systems. There exists some research on attribute reduction in an incremental manner based on rough set theory. Most incremental learning algorithms have been proposed to address dynamic complete decision systems [4,5,25–29,31,32,36]. Incremental attribute reduction has shown its importance in the aspect of efficiency. To our knowledge, previous work on incremental attribute reduction has mainly been concerned with the situation where a single object immigrates into or emigrates from a complete decision system [25–27,29]. However, with the volume of data growing at an unprecedented speed, multiple objects may immigrate into or emigrate from an incomplete decision system simultaneously. Thus, in this paper, an incremental attribute reduction algorithm is developed for computing a new attribute reduct for the case where multiple objects change dynamically in incomplete decision systems. Moreover, in view of the attribute reduction algorithms under dynamic incomplete decision systems, which have not been discussed thus far, we also propose an incremental attribute reduction algorithm for the immigration and emigration of a single object in incomplete decision systems.

The main contributions include: (1) an incremental attribute reduction algorithm is developed for the immigration and emigration of a single object, rather than to obtain the dynamic system as a new one that has to be computed from scratch. (2) Another incremental attribute reduction algorithm is proposed for the immigration and emigration of multiple objects, rather than to perform the reduction algorithm repeatedly. (3) The efficiency and effectiveness of the proposed algorithms are demonstrated on different data sets.

The rest of the paper is organized as follows. Section 2 reviews some related work on attribute reduction in incomplete decision systems and incremental learning techniques in rough set theory. Section 3 introduces some preliminaries on rough set theory and relevant concepts involved in the paper. Section 4 presents the classical attribute reduction algorithm based on the positive region in incomplete decision systems. In Section 5 and Section 6, incremental attribute reduction algorithms are developed in incomplete decision systems for the immigration and emigration of a single object and multiple objects, respectively. In Section 7, the experiments are conducted to evaluate the performance of the proposed algorithms. Section 8 presents conclusions and outlines our further research trends.

2. Related work

Here, we briefly review some representative methods of attribute reduction in the context of incomplete decision systems. From the viewpoint of a discernibility matrix, Kryszkiewicz [8] used a generalized discernibility matrix to obtain all of the attribute reducts

Please cite this article as: W. Shu, W. Qian, An incremental approach to attribute reduction from dynamic incomplete decision systems in rough set theory, Data Knowl. Eng. (2015), http://dx.doi.org/10.1016/j.datak.2015.06.009