International Journal of Approximate Reasoning 76 (2016) 80–95

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

An incremental attribute reduction approach based on knowledge granularity under the attribute generalization

Yunge Jing [a,b], Tianrui Li [a,*], Junfu Huang [a], Yingying Zhang [a]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] Department of Public Computer Teaching, Yuncheng University, Yuncheng 044000, China

ARTICLE INFO

Article history:
Received 17 November 2015
Received in revised form 26 April 2016
Accepted 2 May 2016
Available online 6 May 2016

Keywords:
Decision system
Knowledge granularity
Attribute reduction
Incremental learning
Rough set theory

ABSTRACT

Attribute reduction is a key step to discover interesting patterns in the decision system with numbers of attributes available. In recent years, with the fast development of data processing tools, the information system may increase quickly in attributes over time. How to update attribute reducts efficiently under the attribute generalization becomes an important task in knowledge discovery related tasks since the result of attribute reduction may alter with the increase of attributes. This paper aims for investigation of incremental attribute reduction algorithm based on knowledge granularity in the decision system under the variation of attributes. Incremental mechanisms to calculate the new knowledge granularity are first introduced. Then, the corresponding incremental algorithms are presented for attribute reduction based on the calculated knowledge granularity when multiple attributes are added to the decision system. Finally, experiments performed on UCI data sets and the complexity analysis show that the proposed incremental methods are effective and efficient to update attribute reducts with the increase of attributes.

1. Introduction

Attribute reduction has attracted much attention recently, as an important data preprocessing tool to improve recognition accuracy and discover potentially useful knowledge in many research areas such as knowledge discovery, pattern recognition, expert system, data mining, decision supporting and machine learning [1–9]. In practice, due to many real data sets may increase dynamically in attributes nowadays. Non-incremental approaches are often infeasible since they need to compute repeatedly and consume a large amount of computational time, while incremental approaches are considered as effective techniques to deal with such data because they can directly run the computation using the previous results from the original decision system.

Attribute reduction for a data set based on Rough Set Theory (RST) has many successful applications since it may keep the same information of the original decision system. Many heuristic attribute reduction approaches have been developed based on information entropy, positive region, discernibility matrix, decision cost and knowledge granularity [10–15]. However, these methods are only effectively applied in static decision systems and very inefficient to deal with dynamic decision systems. In practice, there are a lot of examples about dynamic variation of the attribute set in many aspects such as risk prediction, image processing, etc. For example, in a distributive decision system, we need to centralize all data from dif-

* Corresponding author.
E-mail addresses: jyg701022@163.com (Y. Jing), trli@swjtu.edu.cn (T. Li), 990504422@qq.com (J. Huang), yingzhang86@126.com (Y. Zhang).

http://dx.doi.org/10.1016/j.ijar.2016.05.001

ferent locations, which may result in the expansion of attributes in the decision system. Therefore, it needs to develop an incremental attribute reduction algorithm for dynamically varying data sets to obtain valuable knowledge efficiently.

In view of dynamically increasing information systems, many incremental updating algorithms on the variation of the attribute set have been proposed [16–20,33]. Based on upper and lower boundary sets, Chan proposed an incremental algorithm to calculate approximations in RST when one attribute is added to or deleted from the information system [21]. Li et al. developed an incremental approach to compute approximations under the attribute generalization in incomplete information systems [22]. For the incomplete decision system, Shu et al. proposed a positive region-based method for updating the attribute reduct efficiently with the dynamically varying attribute set [23]. Cheng constructed two incremental methods for fast updating approximations through boundary sets and cut sets based on rough fuzzy sets, respectively [25]. Wang et al. proposed a dimension incremental strategy for updating attribute reduct based on information entropy, then proposed an algorithm which is efficient to find a new reduct with dynamically increasing attributes in decision systems [26]. Zeng et al. presented incremental approaches to update attribute reduct of HIS based on fuzzy rough sets when some attributes are added to or deleted from the information system [27]. Li et al. constructed a dominance matrix to compute dominating and dominated sets when the attribute set varies. Then they proposed an incremental method for computing approximations [28]. In set-valued information systems, Luo et al. introduced an incremental mechanism for updating relevant matrices, and developed incremental algorithms for calculating approximations [29].

Matrix is a very useful computing tool for dealing with the decision system. Its theories in calculation and methods are important and they have been indispensable in modern physical science, economics, biology and computer science. Many matrix-based incremental learning algorithms have been proposed to deal with dynamic data sets [24,34,35,39–41]. Most of them mainly focused on updating approximations when the decision system varies dynamically. To fully explore properties in updating reducts, this paper proposes a matrix-based incremental reduction algorithm for dynamic data sets based on the knowledge granularity. It is shown that the matrix-based algorithm for updating the reduct is efficient when the data sets are small. However, for the large data sets, matrix-based incremental algorithm is inefficient since it needs more memory and computational time. To overcome this deficiency, an efficient incremental reduction algorithm based on non-matrices is developed. Then, a series of experiments is performed on 6 data sets from UCI. Experimental results show that the proposed incremental method based on non-matrix is faster than the matrix-based incremental method. Also its computation time is much smaller than that of the non-incremental counterpart and incremental algorithms based on entropy and positive region.

The remainder of this paper is arranged as follows. Section 2 briefly reviews some basic concepts of RST and knowledge granularity. Section 3 introduces a matrix presentation of the knowledge granularity and a general heuristic reduction algorithm based on knowledge granularity for the decision system. In Section 4, incremental reduction algorithms based on matrix and non-matrix when adding multiple attributes are presented. In Section 5, experiments are performed to verify the efficiency and effectiveness of the proposed algorithms. The paper ends with conclusions and the future research in Section 6.

2. Preliminaries

In this subsection, we review several basic definitions of knowledge granularity in RST [30,31].

Definition 1. [31] Given a decision system $S = (U, C \cup D, V, f)$ and $U/IND(C) = \{X_1, X_2, \cdots, X_m\}$. The knowledge granularity of C is defined as

$$GP_U(C) = \sum_{i=1}^{m} \frac{|X_i|^2}{|U|^2}, \tag{1}$$

where the equivalence relation $IND(C)$ is determined by nonempty subset C as follows: $IND(C) = \{(x, y) \in U \times U | \forall a \in C, f(x,a) = f(y,a)\}$.

Definition 2. [30] Given a decision system $S = (U, C \cup D, V, f)$, $U/IND(C) = \{X_1, X_2, \cdots, X_m\}$ and $U/IND(C \cup D) = \{Y_1, Y_2, \cdots, Y_n\}$. A knowledge granularity of C relative to D is defined as follows

$$GP_U(D|C) = GP_U(C) - GP_U(C \cup D). \tag{2}$$

The relative knowledge granularity was used to construct the heuristic attribute reduction algorithm [30]. This reduction algorithm generates a feature subset that has the same discernibility ability as the original one.

Definition 3. [30] Given a decision system $S = (U, C \cup D, V, f)$, $U/IND(C) = \{X_1, X_2, \cdots, X_m\}$ and $B \subseteq C$. $\forall a \in B$, the significance measure (inner significance) of a in B is defined as

$$Sig_C^{inner}(a, B, D) = GP_U(D|(C - \{a\})) - GP_U(D|C). \tag{3}$$