Knowledge-Based Systems 104 (2016) 24–38

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An incremental approach for attribute reduction based on knowledge granularity

Yunge Jing[a,b], Tianrui Li[a,*], Chuan Luo[c], Shi-Jinn Horng[d], Guoyin Wang[e], Zeng Yu[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
[c] College of Computer Science, Sichuan University, Chengdu 610065, China
[d] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan
[e] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Article history:
Received 4 July 2015
Revised 2 April 2016
Accepted 8 April 2016
Available online 21 April 2016

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

ABSTRACT

Rough set provides a theoretical framework for classification learning in data mining and knowledge discovery. As an important application of rough set, attribute reduction, also called feature selection, aims to reduce the redundant attributes in a given decision system while preserving a particular classification property, e.g., information entropy and knowledge granularity. In view of the dynamic changes of the object set in a decision system, in this paper, we focus on knowledge granularity-based attribute reduction approach when some objects vary dynamically. We first introduce incremental mechanisms to compute new knowledge granularity. Then, the corresponding incremental algorithms for attribute reduction are developed when some objects are added into and deleted from the decision system. Experiments conducted on different data sets from UCI show that the proposed incremental algorithm can achieve better performance than the non-incremental counterpart and incremental algorithm based on entropy.

1. Introduction

Rough Set Theory (RST), as an important soft computing tool to discover the potentially useful knowledge from databases, has attracted much attention. Nowadays, it has become a common technique in many research areas such as machine learning [1], pattern recognition [19], decision supporting [28,29], expert system [20], data mining and knowledge discovery [14,15,17,30].

Attribute reduction is an important concept in RST, which keeps the distinguishing power of original decision system for the objects from the universe. Many methods for attribute reduction have been developed in the last two decades. However, these approaches are often applied in static decision systems [10,14,21–25], which are known as a kind of non-incremental reduction algorithms. As the object set varies dynamically in a given decision system, one needs to recompute the decision system from scratch to obtain a new reduct and thus consume a great deal of computational time. Apparently, these reduction algorithms are inefficient to deal with dynamic decision systems.

To overcome this deficiency, a RST-based incremental learning method is commonly used to deal with dynamic data sets [13,18], which carries out the computation using the existing results from the original data set. And it has been successfully applied to data analysis in real-time applications. Researchers have proposed many incremental algorithms to update approximations, reducts and decision rules in dynamic decision systems, which can be divided into three aspects as follows. 1) Incremental updating algorithms on the change of the attribute set. Based on upper and lower boundary sets, Chan developed an incremental algorithm for updating upper and lower approximations. It is an effective method to deal with dynamic attribute sets [2]. Li et al. proposed an incremental method for fast computing approximations by the characteristic relation when the attribute set varies in incomplete decision systems [8]. Zhang et al. developed a matrix-based incremental algorithm to fast compute approximations under the variation of the attribute set in set-valued information systems [31]. Cheng developed two incremental approaches to compute approximations based on cut sets and boundary sets, respectively [5]. Wang et al. developed a dimension incremental algorithm based on information entropy for updating the attribute reduct, and then proposed an algorithm that can find a new reduct efficiently when an attribute set was added to a decision system [21]. In [7], Li et al. introduced the dominance matrix to compute dominating

* Corresponding author. Tel.: +8602886466426.
E-mail addresses: jyg701022@163.com (Y. Jing), trli@swjtu.edu.cn (T. Li), cluo@scu.edu.cn (C. Luo), horngsj@yahoo.com.tw (S.-J. Horng), wanggy@ieee.org (G. Wang), yuzeng2005@163.com (Z. Yu).

http://dx.doi.org/10.1016/j.knosys.2016.04.007
0950-7051/

and dominated sets, and proposed an incremental method for updating approximations when the attribute set varies. Luo et al. developed an incremental approach to update relevant matrices, and presented incremental methods for computing approximations in set-valued information systems [11]. For incomplete decision systems, Shu et al. developed a positive region-based attribute reduction approach to update the attribute reduct efficiently with dynamic changes of the attribute set [16]. 2) Incremental updating algorithms on the change of the object set. Liu et al. presented an incremental attribute reduction algorithm to compute the minimal reduct in an information system [10]. Orlowska et al. presented an incremental reduction method for updating the attribute reduct in a decision system [13]. Considering that the presented method is very time-consuming, Hu et al. proposed an effective incremental reduction algorithm based on the positive region to incrementally update reducts [6]. Liang et al. developed a group incremental attribute reduction algorithm based on information entropy when adding multiple objects into a decision system [9]. Yang et al. proposed an incremental method to update attribute reduct based on the discernibility matrix [27]. 3) Incremental updating algorithms on the change of the attribute values. Chen et al. introduced the definitions of refining and coarsening, and developed the incremental approaches for computing approximations when coarsening or refining attribute values in single-valued decision systems [3]. Based on three representative entropies, Wang et al. presented an attribute reduction algorithm for data sets with dynamic changes of attribute values [22]. Chen et al. introduced the definition of minimal discernibility attribute set, which focuses on improving the efficiency of attribute reduction. Furthermore, they developed a rough set-based approach for updating decision rules in inconsistent decision systems [4].

The knowledge granularity has become one of the most popular uncertainty measures of data sets. It is commonly used to construct heuristic reduction algorithms [12,24,25]. From the above discussion, most incremental reduction algorithms are based on information entropy or positive region when data sets vary dynamically. To our best knowledge, incremental reduction algorithm based on knowledge granularity has not been discussed so far. In addition, matrix is a very useful tool in computing, its theories in calculation and methods are important and they have been indispensable in economics, biology, and modern physical science. To fully explore properties in updating attribute reducts when some objects vary in the decision systems, this paper aims to design incremental approaches for dynamic data sets based on knowledge granularity. In view of that a key step of the incremental reduction approach is the computation of knowledge granularity, this paper first introduces incremental mechanisms of knowledge granularity and develops a matrix-based incremental attribute reduction approach based on knowledge granularity under the variation of objects. However, matrix-based methods need more space and computational time with the increasing size of data sets in real applications. Then, non-matrix based incremental attribute reduction approaches are proposed when some objects are added into or deleted from the decision systems. Finally, the experimental results verify the proposed incremental reduction algorithms can achieve better performance than the incremental algorithm based on information entropy and the non-incremental counterpart.

The remainder of this paper is organized as follows. Section 2 briefly reviews some basic concepts in RST. Section 3 introduces the representation of the knowledge granularity and a general heuristic reduction algorithm based on knowledge granularity. Incremental reduction algorithms based on matrix and non-matrix are developed when some objects are added into or deleted from the decision system in Section 4. In Section 5, experimental evaluation on nine UCI data sets are outlined to demonstrate the effectiveness and efficiency of the proposed algorithms. The paper ends with conclusions in Section 6.

2. Preliminaries

We review some basic concepts of information system and RST in this section [14].

Definition 1. [14] An information system is a quadruple tuple $S=(U,A,V,f)$, where $U=\{u_1,u_2,\cdots,u_n\}$ is a finite non-empty object set and A is a finite nonempty attribute set, $V=\cup_{a\in A}V_a$, V_a is a domain of attribute a, and f: $U\times A\rightarrow V$ is an information function with $f(x,a)\in V_a$ for each $a\in A$ and $x\in U$. If $A=C\cup D$, where C is the conditional attribute set and D is the decision attribute set, then $S=(U,C\cup D,V,f)$ is also called as a decision system. For simplicity, $U=\{u_1,u_2,\cdots,u_n\}$ is written as U in the following:

Definition 2. [14] An indiscernibility (equivalence) relation is determined by each nonempty subset $B\subseteq A$ as follows:

$$R_B=\{(x,y)\in U\times U|f(x,a)=f(y,a),\forall a\in B\} \tag{1}$$

The universe U is divided into some equivalence classes by the indiscernibility relation R_B, given by $U/R_B=\{[x]_B|x\in U\}$, and the equivalence class including the object x with respect to B is denoted as $[x]_B=\{y\in U|(x,y)\in R_B\}$.

Definition 3. [14] Let $S=(U,C\cup D,V,f)$ be a decision system and $X\subseteq U$. R is an equivalence relation. The lower and upper approximations of X with respect to R are defined as follows, respectively:

$$\underline{R}X=\cup\{x\in U|[x]_R\subseteq X\}, \tag{2}$$

$$\overline{R}X=\cup\{x\in U|[x]_R\cap X\neq\varnothing\}. \tag{3}$$

The universe U is partitioned into three disjoint regions by these two approximations $(\underline{R}X,\overline{R}X)$: the negative region $NEG_R(X)$, the boundary region $BND_R(X)$, and the positive region $POS_R(X)$. Then the three regions are defined as follows, respectively [14]:

$$\begin{cases} NEG_R(X)=U-\overline{R}X\\ BND_R(X)=\overline{R}X-\underline{R}X\\ POS_R(X)=\underline{R}X\end{cases}$$

3. Attribute reduction based on knowledge granularity

This section reviews some basic concepts of knowledge granularity and relation matrix. Then some properties are obtained and a general heuristic attribute reduction algorithm based on knowledge granularity for decision systems is introduced.

3.1. A representation of the knowledge granularity

Definition 4. [12] Let $S=(U,C\cup D,V,f)$ be a decision system and $U/C=\{X_1,X_2,\cdots,X_m\}$. Based on the partition, a knowledge granularity of C is defined as

$$GP_U(C)=\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2} \tag{4}$$

The heuristic attribute reduction algorithms based on knowledge granularity aim to preserve the relative knowledge granularity unchanged. Then the relative knowledge granularity shown as follows was applied to construct heuristic reduction algorithm and delete the redundant attributes in a decision system [14].