Knowledge-Based Systems 64 (2014) 59–69

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Updating multigranulation rough approximations with increasing of granular structures

Xibei Yang [a,b,*], Yong Qi [c], Hualong Yu [a], Xiaoning Song [a], Jingyu Yang [b,d]

[a] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, PR China
[b] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Nanjing University of Science and Technology, Ministry of Education, Nanjing, Jiangsu 210094, PR China
[c] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, PR China
[d] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, PR China

ARTICLE INFO

Article history:
Received 3 September 2012
Received in revised form 28 November 2013
Accepted 25 March 2014
Available online 2 April 2014

Keywords:
Attribute reduction
Granular structure
Lower approximation
Multigranulation rough set
Upper approximation

ABSTRACT

Dynamic updating of the rough approximations is a critical factor for the success of the rough set theory since data is growing at an unprecedented rate in the information-explosion era. Though many updating schemes have been proposed to study such problem, few of them were carried out in a multigranulation environment. To fill such gap, the updating of the multigranulation rough approximations is firstly explored in this paper. Both naive and fast algorithms are presented for updating the multigranulation rough approximations with the increasing of the granular structures. Different from the naive algorithm, the fast algorithm is designed based on the monotonic property of the multigranulation rough approximations. Experiments on six microarray data sets show us that the fast algorithm can effectively reduce the computational time in comparison with the naive algorithm when facing high dimensional data sets. Moreover, it is also shown that fast algorithm is useful in decreasing the computational time of finding both traditional reduct and attribute clustering based reduct.

1. Introduction

Rough set theory [20,21,32] is a mathematical tool for characterizing the uncertainty by the difference between the lower and upper approximations. Presently, rough set has been demonstrated to be useful in decision making [18], feature selection [3,8], clustering analysis [9], machine learning [30–33] and so on. A fundamental concept in Pawlak's rough set model is the indiscernibility relation, which is an equivalence relation. By such relation, the equivalence classes are then regarded as the basic knowledge for the construction of the lower and upper approximations. The computation of approximations is a necessary step for attribute reduction and knowledge discovery.

It should be noticed that the volume of the data is growing at an unprecedented rate in many real-world applications and then how to acquire knowledge from data sets is an expensive operation [44]. This phenomenon occurs in several fields, including financial analysis, population study and medical research. This is why many researchers have paid their great attentions to dynamic variation of the data sets. From the viewpoint of the rough set theory, the following dynamic variations have been considered.

1. Dynamic variation of the conditional attributes. For example, Chan [2] presented a theoretical result on updating Pawlak's rough approximations; Li et al. [13] proposed an incremental approach to update approximations under the characteristic relation-based rough set; Zhang et al. [43] compared the non-incremental and incremental approaches to the computations of rough approximations in set-valued information systems; Cheng and Miao [4] presented two incremental methods for fast computing the rough fuzzy approximations.
2. Dynamic variation of the objects. For example, Liu et al. [16] designed an incremental algorithm for inducing knowledge when the universe varies over time; Zhang et al. [44] studied the non-incremental and incremental approaches to update the neighborhood-based rough approximations; Liang et al. [14] developed a group incremental rough feature selection algorithm when a group of the objects are added into a decision table.

* Corresponding author at: No. 2, Mengxi Road, School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, PR China. Tel.: +86 0511 84433445.

E-mail addresses: yangxibei@hotmail.com (X. Yang), qyong@njust.edu.cn (Y. Qi), yuhualong@just.edu.cn (H. Yu), xnsong@yahoo.com.cn (X. Song), yangjy@mail.njust.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.knosys.2014.03.021
0950-7051/© 2014 Elsevier B.V. All rights reserved.

3. Dynamic variation of the attribute values. For example, Chen et al. [5–7] proposed an incremental algorithm for updating approximations under coarsening or refining of attribute values; Wang et al. [34] developed an attribute reduction algorithm for data sets with dynamically varying values.

Obviously, the above approaches are all around the rough set, which is constructed on the basis of one and only one binary relation, and then we call it the single granulation rough set since one binary relation induces one *information granulation* on the universe of discourse. Nevertheless, it should be noticed that in Ref. [22], the authors said that we often need to describe concurrently a target from some independent environments, that is, multiple information granulations are important for problem solving. For example, to extract decision rules from distributive information systems and groups of intelligent agents through using rough set approach, to improve the efficiency of feature selection [24,25] and so on. Therefore, Qian et al. [22,23,26,28] proposed the concept of the multigranulation rough set. The main difference between Pawlak's rough set and multigranulation rough set is that we can use a set of the information granulations for the construction of approximations. Such set of the information granulations can be derived from a family of the binary relations [19,38,40,41], coverings [37], neighborhoods [15], samples, knowledge bases and so on. Moreover, from the viewpoint of the Granular Computing, an information granulation is called a *granular structure* [27,42] (in some rough set literatures, it is also referred to as a partition space [1]). Therefore, following Qian et al.'s multigranulation rough set theory, a set of the information granulations is referred to as a *multigranular structure* in the context of this paper.

From discussions above, the purpose of this paper is to explore the updating computation of the multigranulation rough approximations. Such problem comes from the practice of everyday life. For instance, in a distributive information system, new information may be obtained if new sensors are installed; in venture-investment decision making, new decision may be made if new experts are invited to make evaluations. In our updating of multigranulation rough approximations, both naive and fast algorithms are presented. For naive algorithm, all objects and granular structures should be re-scanned while for fast algorithm, previous results are fully used to decrease the computational time.

To facilitate our discussions, we present the basic knowledge about Pawlak's rough set and two typical multigranulation rough sets in Section 2. In Section 3, we develop naive and fast approaches to update multigranulation rough approximations with the increasing of the granular structures. In Section 4, the performances of naive and fast approaches are evaluated on six microarray data sets with high dimensional genes. Furthermore, to show that the proposed approach is useful in solving real-world problems, the fast scheme is used to find traditional reduct and attribute clustering based reduct. Experimental results show that fast approach can reduce the time significantly. The paper ends with conclusions and outlooks for further research in Section 5.

2. Preliminary knowledge on rough sets

2.1. Rough set

Formally, an information system I can be considered as a pair $\langle U, AT\rangle$, in which U is a non-empty finite set of the objects called the universe, AT is a non-empty finite set of the attributes. $\forall a \in AT, V_a$ is the domain of attribute a. $\forall x \in U, a(x)$ denotes the value that x holds on a ($\forall a \in AT$). Given an information system $I, \forall A \subseteq AT$, an indiscernibility relation $IND(A)$ may be defined as

$$IND(A) = \{(x,y) \in U^2 : a(x) = a(y),\ \forall a \in A\}. \tag{1}$$

Obviously, $IND(A)$ is an equivalence relation. By indiscernibility relation $IND(A)$, $\forall X \subseteq U$, one can construct the lower and upper approximations of X such that

$$\underline{A}(X) = \{x \in U : [x]_A \subseteq X\} \quad \text{and} \quad \overline{A}(X) = \{x \in U : [x]_A \cap X \neq \emptyset\}, \tag{2}$$

respectively, where $[x]_A = \{y \in U : (x,y) \in IND(A)\}$ is the equivalence class of x in terms of set of the attributes A. The partition $U/IND(A) = \{[x]_A : x \in U\}$ is a granular structure induced by A. The pair $[\underline{A}(X), \overline{A}(X)]$ is referred to as a Pawlak's rough set of X with respect to A.

2.2. Multigranulation rough sets

Qian et al.'s multigranulation rough set approach is different from Pawlak's rough set approach since the former is constructed on the basis of a family of the binary relations instead of a single one binary relation. In the following, to simplify our discussions, it is assumed that each attribute in an information system I is corresponding to an equivalence relation. In other words, each attribute in I can induce a granular structure on the universe of discourse and then all the attributes in I will induce a multigranular structure.

In Qian et al.'s multigranulation rough set theory, two different models have been defined. The first one is the optimistic multigranulation rough set, the second one is the pessimistic multigranulation rough set. For lower approximation, the word "optimistic" is used to express the idea that in multigranular structure, we need at least one of the granular structures to satisfy the inclusion condition. The optimistic multigranulation upper approximation is defined by the complement of the optimistic multigranulation lower approximation.

Definition 1 [28]. Let I be an information system, in which $a_1, a_2, \ldots, a_m \in AT$, then $\forall X \subseteq U$, the optimistic multigranulation lower and upper approximations of X are denoted by $\underline{\sum_{i=1}^{m} a_i{}^O}(X)$ and $\overline{\sum_{i=1}^{m} a_i{}^O}(X)$, respectively,

$$\underline{\sum_{i=1}^{m} a_i{}^O}(X) = \{x \in U : [x]_{a_1} \subseteq X \vee \cdots \vee [x]_{a_m} \subseteq X\}; \tag{3}$$

$$\overline{\sum_{i=1}^{m} a_i{}^O}(X) = \sim \left(\underline{\sum_{i=1}^{m} a_i{}^O}(\sim X) \right); \tag{4}$$

where $[x]_{a_i}$ $(1 \leqslant i \leqslant m)$ is the equivalence class of x in terms of the attribute a_i, $\sim X$ is the complement of set X.

Proposition 1 [36]. *Let I be an information system, in which $a_1, a_2, \ldots, a_m \in AT$, then $\forall X \subseteq U$, we have*

$$\overline{\sum_{i=1}^{m} a_i{}^O}(X) = \{x \in U : [x]_{a_1} \cap X \neq \emptyset \wedge \cdots \wedge [x]_{a_m} \cap X \neq \emptyset\}. \tag{5}$$

By Proposition 1, we can see that the optimistic multigranulation upper approximation can be considered as a set, in which objects intersect the target for all the granular structures.

The pessimistic multigranulation rough set is different from the optimistic case. For lower approximation, the word "pessimistic" is used to express the idea that we need all the granular structures to satisfy the inclusion condition. The pessimistic multigranulation upper approximation is also defined by the complement of the pessimistic multigranulation lower approximation.