Knowledge-Based Systems 79 (2015) 90–98

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Fast approach to knowledge acquisition in covering information systems using matrix operations

Anhui Tan [a,b], Jinjin Li [a,*], Guoping Lin [a], Yaojin Lin [c]

[a] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, Fujian, China
[b] School of Mathematical Sciences, Xiamen University, Xiamen 361005, Fujian, China
[c] School of Computer Science, Minnan Normal University, Zhangzhou 363000, Fujian, China

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form 15 January 2015
Accepted 1 February 2015
Available online 9 February 2015

Keywords:
Rough set
Covering information system
Knowledge acquisition
Set approximation
Reduct
Matrix

ABSTRACT

Covering rough set theory provides an effective approach to dealing with uncertainty in data analysis. Knowledge acquisition is a main issue in covering rough set theory. However, the original rough set methods are still expensive for this issue in terms of time consumption. To further improvement, we propose fast approaches to knowledge acquisition in covering information systems by employing novel matrix operations. Firstly, several matrix operations are introduced to compute set approximations and reducts of a covering information system. Then, based on the proposed matrix operations, the knowledge acquisition algorithms are designed. In the end, experiments are conducted to illustrate that the new algorithms can dramatically reduce the time consumptions for computing set approximations and reducts of a covering information system, and the larger the scale of a data set is, the better the new algorithms perform.

1. Introduction

Rough set theory, proposed by Pawlak [16], is one of the effective tools for uncertainty management. It has been demonstrated be useful in a wide variety of applications related to knowledge discovery and data mining [9,34]. Classical rough set theory based on equivalence relations is not applicable for various types of data sets such as numerical ones, set-valued ones, missing ones and interval-valued ones. As a result, many generalized rough set models have been developed in terms of different requirements. For example, Hu et al. [11,12] used a neighborhood-based rough set to deal with numerical data sets. Leung et al. [13] and Abo-Tabl [1] defined tolerance and similarity relations on data sets with missing values. Qian and Liang [20,21] constructed multigranulation rough sets from viewpoint of Granular Computing originated by Zadeh [36,37]. As we know, classical rough sets and these generalized models have a limitation for processing real-valued data sets. Tn address this issue, many scholars have combined the ideas of rough sets and fuzzy sets, and developed new theoretical mechanisms for data analysis, such as fuzzy rough sets [8,10], rough fuzzy sets [2,8] and bipolar fuzzy rough sets [28,29]. In fact, the neighborhood relations [23,41], tolerance relations [7,31], similarity relations [6,15] and fuzzy relations [4,27] all generate coverings of the universe. Therefore, these kinds of rough sets can be categorized into the so-called covering rough sets.

Knowledge acquisition is the main issue in covering rough set theory, especially on the construction of set approximations. Żakowski [38] first proposed the rough set model based on coverings. Pomykala [17] developed two definitions of approximations by naturally generalizing that in classical rough sets. Zhu et al. [42,43] introduced several types of approximations and examined three types of covering rough set models under these approximations. It should be pointed out that Yao et al. [33,35] summarized the existing studies on the approximations of covering rough sets by elements, granules and subsystems-based definitions of approximation operators. However, all these works focus on the set approximations and do not concern with the knowledge reduction of covering rough sets. In order to acquire knowledge in covering rough sets more quickly and efficiently, the notion of knowledge reduction is proposed to reduce redundant members under the condition of preserving certain qualities [3,42]. From this viewpoint, the members in a reduct are jointly sufficient and individually necessary [22]. Within authors' knowledge, there are mainly two types of covering reductions in research community. One is to determine a minimal subcovering of a covering that preserves the original approximations. Zhu et al. [42,43] systematically investigated this type of reduction by employing union-reducible elements in a covering. But the reduction is not in line

* Corresponding author.
E-mail addresses: shujujiegouwang@126.com (A. Tan), jinjinli@mnnu.edu.cn (J. Li), guoplin@163.com (G. Lin), yjlin@mail.hfut.edu.cn (Y. Lin).

http://dx.doi.org/10.1016/j.knosys.2015.02.003
0950-7051/

with the original purpose of knowledge reduction in classical rough set theory [3]. Compared with this reduction in a single covering, Chen and Wang [4,25] first defined the notion of covering decision information system and presented a pioneering study on the reduction of this information system. More research has also concentrated on the covering decision information systems. Li and Yin [14] proposed a reduction method in covering decision systems based on information theory. Wang et al. [24] investigated the relationship of reducts between different covering information systems. Zhang et al. [40] proposed a reduction method to extract decision rules in a covering decision system. Yang et al. [30] employed the minimal elements in covering decision systems to reduce redundant coverings. As we know, the computations of concepts in covering information systems are difficult tasks because the data sets usually overlap and their set operations are much time-consuming. On the other hand, databases expand quickly nowadays and the big data analysis is receiving much attention. However, the original rough set methods are still computationally very expensive for knowledge acquisition in covering information systems, especially on large-scale data sets. Thus how to acquire knowledge from overlapping data with a more efficient way is a desirable topic.

In this paper, we wish to develop faster approaches to knowledge acquisition in covering information systems. We introduce several novel matrix operations into covering information systems, by which the time consumptions for computing both the approximations of sets and reducts of covering information systems can be dramatically reduced. Moreover, the larger the scale of data sets is, the better the matrix operations perform.

The study is organized as follows. The covering rough sets and covering information systems are briefly reviewed in Section 2. In Section 3, several matrix representations and matrix operations are introduced to covering information systems, by which the matrix-based methods are constructed to compute the approximations of sets. Moreover, the methods for incrementally updating approximations in a dynamic covering system is considered. In Section 4, a discernibility matrix is developed which can be used to calculate all the reducts and one suboptimal reduct of a covering system. Bases on these discussions, knowledge acquisition algorithms are designed using the proposed matrix operations. In Section 5, several numerical experiments are conducted on UCI data sets and microarray data sets, showing that the proposed methods are much faster than some commonly used ones in terms of computational time.

2. Background

In this section, we review basic concepts related to covering rough sets and covering information systems. More details can be found in [3,35].

Definition 1. Let U be a universe and C be a family of subsets of U. If none subsets in C is empty and $\cup C = U$, then C is called a covering of U. The ordered pair (U, C) is called a covering approximation space.

One can see that a partition of U is certainly a covering of U.

Definition 2 (*Neighborhood*). Let C be a covering of U. For $x \in U$, denote $C_x = \cap\{K \in C | x \in K\}$ as the neighborhood of $x \in U$ w.r.t. C. Then $Cov(C) = \{C_x | x \in U\}$ is also covering of U and we call it the induced covering of C.

Definition 3 (*Neighborhood*). Let Δ be a family of coverings of U. For each $x \in U$, denote $\Delta_x = \cap\{C_x | C \in \Delta\}$ as the neighborhood of $x \in U$ w.r.t. Δ. Then $Cov(\Delta) = \{\Delta_x | x \in U\}$ is also a covering of U and we call it the induced covering of Δ.

Let Δ be a family of coverings of U. The ordered pair $S = (U, \Delta)$ is called a covering information system.

Definition 4. Let $S = (U, \Delta)$ be a covering information system. A pair of approximation operators $(\underline{\Delta}, \overline{\Delta})$ is defined as: $X \subseteq U$,

$\underline{\Delta}(X) = \{x \in U | \Delta_x \subseteq X\}$;

$\overline{\Delta}(X) = \{x \in U | \Delta_x \cap X \neq \emptyset\}$.

In what follows, we introduce the notion of reduct in a covering information system.

Definition 5. Let $S = (U, \Delta)$ be a covering information system. For a subset of coverings $P \subseteq \Delta$, if $Cov(P) = Cov(\Delta)$, then P is called a consistent set of S. Furthermore, if $Cov(P) = Cov(\Delta)$, and $Cov(Q) \neq Cov(\Delta)$ for any $Q \subset P$, then P is called a reduct of S. The member in Δ that belongs to all reducts is denoted as the core of S.

Clearly, if each member in Δ is a partition, the definition of reduct of covering information systems is accordant with that of traditional information systems.

Lemma 1. *Let $S = (U, \Delta)$ be a covering information system. Then $P \subseteq \Delta$ is a reduct of S iff P is a minimal subset satisfying $\Delta_x = P_x$ for all $x \in U$.*

Proof. By Definition 5, it is straightforward. □

The reduct of a covering information system can be characterized as follows.

Property 1. *Let $S = (U, \Delta)$ be a covering information system. For a subset of coverings $P \subseteq \Delta$, the following propositions are equivalent to each other:*

(1) *P is a reduct of S;*
(2) *P is a minimal subset satisfying $\underline{\Delta}(X) = \underline{P}(X)$ for all $X \subseteq U$;*
(3) *P is a minimal subset satisfying $\overline{\Delta}(X) = \overline{P}(X)$ for all $X \subseteq U$.*

Proof. Since $\underline{\Delta}$ and $\overline{\Delta}$ are dual, we only prove (1) $\Longleftrightarrow$ (2).

(1) $\Rightarrow$ (2): By Lemma 1 and Definition 4, it is obvious.

(2) $\Rightarrow$ (1): By Lemma 1, we need to prove that $\Delta_x = P_x$ for all $x \in U$.

Suppose, by contradiction, that $\Delta_{x_0} \neq P_{x_0}$ for some $x_0 \in U$. It is clear that $\Delta_{x_0} \subset P_{x_0}$. Let us take $X = \Delta_{x_0}$. Then $\Delta_{x_0} \subseteq X$ which implies $x_0 \in \underline{\Delta}(X)$. On the other hand, $P_{x_0} \not\subseteq X$ which implies $x_0 \notin \underline{P}(X)$. Hence $\underline{\Delta}(X) \neq \underline{P}(X)$. It is a contradiction. □

Let us employ the following example in literatures [3,25] to illustrate our idea.

Example 1. [3,25]. Suppose $U = \{x_1, x_2, \ldots, x_9\}$ and $\Delta = \{C_1, C_2, C_3, C_4\}$, where

$C_1 = \{\{x_1, x_2, x_4, x_5, x_7, x_8\}, \{x_2, x_3, x_5, x_6, x_8, x_9\}\}$,

$C_2 = \{\{x_1, x_2, x_3, x_4, x_5, x_6\}, \{x_4, x_5, x_6, x_7, x_8, x_9\}\}$,

$C_3 = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6, x_7, x_8, x_9\}\}$,

$C_4 = \{\{x_1, x_2, x_4, x_5\}, \{x_2, x_3, x_5, x_6\}, \{x_4, x_5, x_7, x_8\}, \{x_5, x_6, x_8, x_9\}\}$.

We can calculate that:

$$C_{1x_1} = \{x_1, x_2, x_4, x_5, x_7, x_8\},\ C_{1x_2} = \{x_2, x_5, x_8\},$$

$$C_{1x_3} = \{x_2, x_3, x_5, x_6, x_8, x_9\},\ C_{1x_4} = \{x_1, x_2, x_4, x_5, x_7, x_8\},$$

$$C_{1x_5} = \{x_2, x_5, x_8\},\ C_{1x_6} = \{x_2, x_3, x_5, x_6, x_8, x_9\},$$

$$C_{1x_7} = \{x_1, x_2, x_4, x_5, x_7, x_8\},\ C_{1x_8} = \{x_2, x_5, x_8\},$$

$$C_{1x_9} = \{x_2, x_3, x_5, x_6, x_8, x_9\}.$$