International Journal of Approximate Reasoning 75 (2016) 13–38

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Data volume reduction in covering approximation spaces with respect to twenty-two types of covering based rough sets

A. Shakiba [a,b], M.R. Hooshmandasl [a,b,*]

[a] Department of Computer Science, Yazd University, Yazd, Iran
[b] Laboratory of Quantum Information Processing, Yazd University, Yazd, Iran

ARTICLE INFO

Article history:
Received 2 July 2015
Received in revised form 2 February 2016
Accepted 9 April 2016
Available online 13 April 2016

Keywords:
Data volume reduction
Covering based rough set
Consistent function
Homomorphism
Approximation space

ABSTRACT

In this paper, we investigate whether consistent mappings can be used as homomorphism mappings between a covering based approximation space and its image with respect to twenty-two pairs of covering upper and lower approximation operators. We also consider the problem of constructing such mappings and minimizing them. In addition, we investigate the problem of reducing the data volume using consistent mappings as well as the maximum amount of their compressibility. We also apply our algorithms against several datasets.

1. Introduction

In today's real life situations, we are generally faced with massive volumes of data in information systems. Hence, a capability in reducing this volume of data is of great importance. In addition to Big data problem, representing the knowledge, as a critical task in the field of artificial intelligence, becomes difficult when the knowledge is imprecise or uncertain. Coping with Big and imprecise knowledge at the same time, doubles the problems.

To handle the imprecise or uncertain data, researchers proposed several methods like rough sets and its extensions [1–4], Dempster–Shafer theory [5], fuzzy sets [6,7], granular computing [8–10] and S-approximation spaces [11–13]. In this paper, we will concentrate on a special extension of rough set called covering rough sets or covering approximation spaces [14]. In covering approximation spaces, there are several approximation operators like [14–17,4,18–29] with different constructions and properties. The concept of covering rough sets is a well studied topic, for example it is studied in a multigranular context [30] and it is shown that covering approximation spaces with certain approximation operators are accordant to formal concepts [31]. Among these studies, there is a line of research for communication or homomorphism among information systems, like [32–34]. Note that the notion of homomorphism was introduced in [35] to study the relation between two information systems in order to obtain a smaller information system where their description functions are identical. Later, other invariants of information systems with respect to different classes of homomorphism were considered by several scholars for different cases, such as [36–39,32,40–42], to mention a few. Up to the knowledge of the authors, there are no studies on homomorphism and its relation to reduction of data volume, except [22] which discussed the problem of data volume reduction in the sense of removing unnecessary elements of a covering. All of these papers concentrate on reducing the dimensionality of the data. In other words, in an incomplete information system like $\mathcal{I} = (U, A, V \cup \{*\}, f)$ [43], the

* Corresponding author at: Department of Computer Science, Yazd University, Yazd, Iran.
E-mail addresses: ali.shakiba@stu.yazd.ac.ir (A. Shakiba), hooshmandasl@yazd.ac.ir (M.R. Hooshmandasl).

http://dx.doi.org/10.1016/j.ijar.2016.04.002
0888-613X/

dimension of the data is $|A|$, whereas the volume size of the data is $|U|$ where U denotes the set of elements or rows and A denotes the set of attributes or columns. For example, in [33] a kind of homomorphism was employed to reduce the dimensionality of data.

In general, to handle large volumes of data, there exist several techniques such as data summarization or approximating the solution using a relatively smaller subset of data [44]. The problem of reducing data volume can be seen as construction of a many to one mapping, i.e. there is one representative element for a set of elements. However, all of such mappings are not necessarily useful. Therefore, in this paper we will investigate a class of mappings for which some special properties of the original covering approximation spaces are preserved in the image with respect to that mapping like definability of sets. In other words, we will study the problem of data volume reduction in covering approximation spaces with respect to 22 pairs of covering lower and upper approximation operators. Along this path, we will show that the notion of consistent functions introduced in [33] can be used to obtain a homomorphism with respect to these types of approximation operators and compress the data by reducing its volume. Up to the authors' knowledge, there are no results for the maximum amount of compressibility of covering spaces by use of consistent functions or an algorithmic approach for their construction. So, this paper also considers this problem in part by introducing the notion of consistency number of a covering.

The organization of this paper is as follows: In Section 2, we will review Pawlak's original formulation of rough set along with twenty-two pairs of covering based approximation operators as well as some relations among them. We will also review basics of consistent mappings. Then, in Section 3, we define the notions of isomorphism and homomorphism with respect to covering approximation operators. In Section 4, we do a detailed investigation of consistent functions and their ability as homomorphism and isomorphism with respect to these operators. In Section 5, we will investigate the process of constructing consistent mappings with respect to a given covering. Also, we will investigate the maximum amount of compressibility of a covering based approximation space using consistent mappings in terms of their consistency numbers and then give an illustrative example as well as applying our algorithms to some datasets. Finally, we conclude the paper and give some future research directions.

2. Preliminaries

In this section, we review concepts like basics of rough set theory, covering based approximation spaces and consistent mappings.

2.1. Rough set theory

Let U be a finite non-empty set, R an equivalence relation on U and $X \subseteq U$. The *lower and upper approximations* of set X with respect to R are defined as

$$\underline{R}(X) = \{Y_i \in U/R | Y_i \subseteq X\},$$

and

$$\overline{R}(X) = \{Y_i \in U/R | Y_i \cap X \neq \emptyset\},$$

where $U/R = \{Y_1, \dots Y_m\}$ is the quotient space of U with respect to R [1]. A set X is called R-*definable* if $\underline{R}(X) = \overline{R}(X) = X$, otherwise it is called R-*undefinable*. We also call an R-definable set a *crisp set* and an R-undefinable set a *rough set*. The concept of rough set is extended and studied in at least four ways: one by considering other kinds of relations on U, e.g. tolerance relations [45], another by consideration of coverings [14], the other by generalizing the relation over two sets [46] and finally by considering different decider functions [13,11,47].

The original Pawlak's rough set was generalized to covering based approximation spaces by W. Zakowski [14] by considering a covering instead of a partition. Since this paper is considered with covering based approximation spaces, we will give an overview of some necessary facts about them.

Definition 2.1 *(Covering and covering approximation space).* (See [22].) Let U be a finite non-empty set and $\mathbf{C}$ a family of non-empty subsets of U such that $\bigcup_{K \in \mathbf{C}} K = U$. Then, $\mathbf{C}$ is called a *covering of* U and the ordered pair $(U, \mathbf{C})$ is called a *covering approximation space*.

Definition 2.2 *($\mathbf{C}$-definable and $\mathbf{C}$-undefinable sets).* (See [22].) Let $(U, \mathbf{C})$ be a covering approximation space and $X \subseteq U$. A set X is called $\mathbf{C}$-*definable* if it can be written as a union of some $K \in \mathbf{C}$. Otherwise, X is called a $\mathbf{C}$-*undefinable* set.

Remark 1. Note that there exists another definition for $\mathbf{C}$-definable sets similar to the one in Pawlak's original formulation, i.e. a set $X \subseteq U$ is called a $\mathbf{C}$-definable set whenever its lower and upper approximations are equal. Although these two definitions are equivalent with respect to some covering lower and upper approximation pairs; but there are pairs which they are not. In this paper, we will stick to the Definition 2.2.

In a covering like $\mathbf{C}$ over U, an element $x \in U$ might belong to more than one $K \in \mathbf{C}$. So, we need some kind of minimal and maximal descriptions.