International Journal of Approximate Reasoning 71 (2016) 89–111

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Compound approximation spaces for relational data

Piotr Hońko

Faculty of Computer Science, Bialystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form 19 December 2015
Accepted 5 February 2016
Available online 9 February 2016

Keywords:
Rough sets
Granular computing
Data mining
Relational databases

ABSTRACT

Rough set theory provides a powerful tool for dealing with uncertainty in data. Application of variety of rough set models to mining data stored in a single table has been widely studied. However, analysis of data stored in a relational structure using rough sets is still an extensive research area. This paper proposes compound approximation spaces and their constrained versions that are intended for handling uncertainty in relational data. The proposed spaces are expansions of tolerance approximation ones to a relational case. Compared with compound approximation spaces, the constrained version enables to derive new knowledge from relational data. The proposed approach can improve mining relational data that is uncertain, incomplete, or inconsistent.

1. Introduction

Dealing with uncertainty in data is a challenging task in the field of data mining. A powerful framework intended for this issue is provided by rough set theory [13]. A concept that can include uncertain data is characterized in this theory by a pair of two certain sets, i.e. its lower and upper approximations. New knowledge about the concept can be derived from the approximations. For example, decision rules constructed based on the lower approximation show features of objects that certainly belong to the concept, whereas those generated from the upper one describe objects whose membership in the concept is possible.

Rough set theory is often considered in the context of granular computing [1,14]. This field can be viewed as a label of theories, methodologies, techniques, and tools that make use of granules in the process of problem solving [30]. Granules are, in general, understood as collections formed in the process of a semantically meaningful grouping of elements based on their indistinguishability, similarity, proximity or functionality [2]. In rough set theory, a granulation is obtained by defining an indiscernibility or similarity relation over the universe of discourse. Therefore, each class of the relation can be seen as an elementary granule.

Many different rough sets models have been proposed over the last three decades (e.g. [36,20,5,29]). They have found wide range of applications in areas such as e.g. medicine, banking, or engineering (for more details, see, e.g. [23,19,15]). The standard rough set model has been generalized in a variety of ways. However, most of the rough set approaches are intended to analyze data stored in a single table. Such a data structure makes it possible to encode simple features of objects of the interest. To show more complex properties such as relationships among objects, a more advanced structure is needed, e.g. relational database.

The following two subsections review works related to the problem of generalization of rough set models.

E-mail address: p.honko@pb.edu.pl.

http://dx.doi.org/10.1016/j.ijar.2016.02.002
0888-613X/

1.1. Generalization of rough set models to multiple universes

In [27] the approximation space is defined as a triple of two distinct universes and binary relation which is a subset of the Cartesian product of the universes. Approximations are defined for a subset of one of the universes. They include objects from the other universe that are in the relation with objects of the subset. Such an approach can be viewed as a generalization of that introduced in [28] where approximations are defined in a formal context that is a triple of a universe of objects, universe of attributes, and a binary relation between the universes.

In [10] approximations are defined in an information system that is a pair of the double universe (the Cartesian product of two particular universes) and the attribute set. Approximations of a subset of the double universe are defined based on equivalence classes of the equivalence relation on the double universe. Additionally, a constrained version of the information system is introduced. It is a triple of the double universe, a constraint relation on the universe, and the attribute set.

To handle with data stored in many tables a multi-table information system is proposed in [12]. The system is a finite set of tables (each table is viewed as an information system). Approximations are defined for a subset of the universe of one specified table, that is, the decision table. Elementary sets of a given universe are used to define the approximations. Indiscernibility of objects from the decision table is defined using the information available in all the tables of the multi-table information system.

An information system for processing data distributed over multiple universes is proposed in [22]. The information system, called a sum of information systems, is the pair of the universe (the Cartesian product of the universes of the information systems, each corresponding to one table) and the attribute set (the collection of attributes from the attribute sets of the information systems). A constrained version of this system allows only tuples of objects that belong to a constraint relation on the Cartesian product of the universes. The constraint relation can be constructed by conditions expressed by Boolean combination of descriptors of attributes. Not only the attributes from the attribute set, but also some other ones specifying relation between particular information systems can be used to define the constraints. Approximation spaces for multiple universes are constructed based on (constraint) sums of information systems. Approximations are defined for a subset of the Cartesian product of the universes using approximations computed for particular information systems.

The tolerance rough set model was adapted in [6] for mining relational data. The universe in this model corresponds the target table of a given database and is defined as the set of granules derived from relational data. Each granule corresponds to one object and includes its description constructed using information stored in the whole database. Approximations are defined analogously to those from the original tolerance rough set model.

1.2. Other generalization of rough set models

The standard rough set model was extended to a covering generalized rough set model (e.g. [3,35]), where the universe is replaced with its covering. Such a generalization enables to deal with more complex problems. Covering rough set theory with the concept of neighborhood induced by covering plays an important role in reduction of nominal data and in generation of decision rules from incomplete data.

In [11] a relationship between different approximation operators defined in covering rough set theory was studied. It was shown that the operators that use the notion of neighborhood and the complementary neighborhood can be defined almost in the same way. It was also investigated that such twin approximation operators have similar properties.

In [25] matrix-based methods for computing approximations of a given concept in a covering decision system is proposed. The methods are also used for reducing covering decision systems. It was shown that the proposed approach can decrease the computational complexity for finding all reducts.

Another generalization of the standard rough set model (single granulation rough set model) is rough set model based on multi-granulations (MGRS) [17]. Approximations of a concept are defined by using multiple equivalence relations on the universe. The relations are chosen according to user requirements or the problem to be solved. MGRS is considered in two different versions. If the condition of the lower approximation is satisfied for (at least one of/all) single granulation rough set models under consideration, then MGRS is called (optimistic/pessimistic). Properties of optimistic and pessimistic multi-granulation rough set models investigated in [18] show connections of these models with notions such as lattices, topology on the universe, and Boolean algebra.

A model that can be viewed as a multi-granulations form of nearness approximation space was introduced in [26]. A topological neighborhood based on $*EI$ algebra (a notion from axiomatic fuzzy set theory) is used in information systems with many category features. The neighborhood is combined with generalized approximation spaces producing, thereby, an extension model of the approximation space.

Another kind of extension of the standard rough set model is composite rough set model [31] that is intended for dealing with multiple types of data in information systems, e.g., categorical data, numerical data, set-valued data, interval-valued data and missing data. All basic rough set notions such as lower and upper approximations, positive, boundary and negative regions are redefined in composite information systems.

A dynamic version of the composite rough set model [32], which uses a matrix-based method, makes it possible to fast update approximations of a changing concept.