Rough-Granular Computing Based Relational Data Mining

Piotr Hońko

Department of Computer Science, Białystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland
p.honko@pb.edu.pl

Abstract. In this paper, we propose a rough-granular computing framework for mining relational data. We adapt the tolerance rough set model for relational data analysis. We introduce two ways for constructing the universe from relational data. Due to applying granular computing methods, one can overcome problems such as relational data representation and the search space limitation. We also show how the proposed framework can be applied to data mining tasks such as classification.

Keywords: multi-relational data mining, rough sets, granular computing, classification.

1 Introduction

Rough set theory [13] as a useful tool to deal with imprecise data is often considered as one of basic techniques of granular computing [2,14]. In this view, granules are formed by means of rough inclusions as classes of objects close to a specified center of the granule to a given degree. Formally, they resemble neighborhoods formed with respect to a certain metric.

In recent years, one can observe a trend in data mining towards the application of granular computing based on the rough set approach. This newly emerging approach is called rough-granular computing [18,15].

Techniques of granular computing, especially rough sets, have widely been applied in the field of data mining (see, e.g., [14,2,15]). Methods of rough sets have also found application in mining data stored in multiple tables, i.e., relational data mining.

(Multi-)relation data mining (MRDM) ([5,4]) concerns knowledge discovery from relational databases consisting of multiple relations (tables). MRDM aims to integrate methods from existing fields applied to an analysis of data represented by multiple relations; producing new techniques for mining multi-relational data.

In MRDM, rough set theory has found application in tasks such as eliminating unimportant data (see, e.g., [17]); the analysis of invalid, missing, and indistinguishable data (see, e.g., [9,11]); reducing data size (see, e.g., [10]); relational classification rules generation (see, e.g., [10,19,12]).

S. Greco et al. (Eds.): IPMU 2012, Part I, CCIS 297, pp. 290–299, 2012.

Despite rough set theory has successfully been adapted in MRDM, none of the current approaches have been defined in a framework of granular computing. Such an approach can be useful in resolving important relational data mining problems, such as adjusting relational data representation to a given task and limiting the search space for relational patterns, e.g., classification rules.

In this paper, we propose a rough-granular computing framework for mining relational data. We adapt the tolerance rough set model [16,17] for mining relational data. We introduce two ways for constructing the universe from relational data: the universe constructed from granules directly derived from relational data and the one constructed from information granules being a generalized representation of relational data. Our framework combines advantages of both granular computing and rough sets. Due to applying granular computing methods, one can overcome the problems of relational data representation and the search space limitation. The application of rough sets makes it possible to effectively analyze relational data for data mining tasks such as classification.

In the following sections, we briefly discuss granular computing and rough sets (Sect. 2), propose a rough-granular computing approach for mining relational data, introduce a tolerance rough set model for relational data, show an application of our approach (Sect. 3), and finally provide conclusion remarks (Sect. 4).

2 Granular Computing and Rough Sets

Granular computing (GC) can be viewed as a label of theories, methodologies, techniques, and tools that make use of granules in the process of problem solving [20]. A granule is a collection of entities drawn together by indistinguishability, similarity, proximity or functionality [21]. The process of the formation of granules is called granulation. To clearly differentiate granulation from clustering, the semantic aspect of granular computing is taken into account. Namely, information granulation is understood as a semantically meaningful grouping of elements based on their indistinguishability, similarity, proximity or functionality [3].

One can obtain many granularities of the same universe, differing their levels. A granule of high-level granularity, i.e., a high-level granule represents a more abstract concept, and a low-level granule a more specific one. A basic task of GC is to switch between different levels of granularity. A more specific level granularity may reveal more detailed information. On the other hand, a more abstract level granularity may improve a problem's solution thanks to omitting irrelevant details.

Rough sets [13] as a useful tool to deal with unprecise data are often considered as one of basic techniques of granular computing. Since Pawlak proposed his rough set model, a number of extensions and generalizations of the model have been introduced. In this work, we use the tolerance rough set model [16,17] due to its flexibility in tuning parameters.