Information Sciences 287 (2014) 68–89

Contents lists available at ScienceDirect

Information Sciences

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

Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets"

Lala Septem Riza [a,*], Andrzej Janusz [b], Christoph Bergmeir [a], Chris Cornelis [a], Francisco Herrera [a], Dominik Ślęzak [b], José Manuel Benítez [a]

[a] *Department of Computer Science and Artificial Intelligence, CITIC-UGR, IMUDS, University of Granada, Spain*
[b] *Institute of Mathematics, University of Warsaw, Poland*

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 12 June 2014
Accepted 24 July 2014
Available online 1 August 2014

Keywords:
Rough set
Fuzzy rough set
Instance selection
Discretization
Feature selection
Rule induction

ABSTRACT

The package *RoughSets*, written mainly in the R language, provides implementations of methods from the rough set theory (RST) and fuzzy rough set theory (FRST) for data modeling and analysis. It considers not only fundamental concepts (e.g., indiscernibility relations, lower/upper approximations, etc.), but also their applications in many tasks: discretization, feature selection, instance selection, rule induction, and nearest neighbor-based classifiers. The package architecture and examples are presented in order to introduce it to researchers and practitioners. Researchers can build new models by defining custom functions as parameters, and practitioners are able to perform analysis and prediction of their data using available algorithms. Additionally, we provide a review and comparison of well-known software packages. Overall, our package should be considered as an alternative software library for analyzing data based on RST and FRST.

1. Introduction

Rough set theory (RST) was introduced by Pawlak in 1982 [75] as a methodology for data analysis based on the approximation of concepts in information systems. It revolves around the notion of discernibility: the ability to distinguish between objects, based on their attribute values. Given an indiscernibility relation, we can construct lower and upper approximations of concepts. Objects included in the lower approximation can be classified with certainty as members of the concept. In contrast, the upper approximation contains objects possibly belonging to the concept. For more than three decades RST has been attracting researchers and practitioners in many different areas. For example, RST is applied in diverse domains such as water quality analysis [52], intrusion detection [63], bioinformatics [57], and character pattern recognition [60]. An important advantage of RST is that it does not require additional parameters to analyze the data.

RST has been generalized in many ways to tackle various problems. In particular, in 1990, Dubois and Prade [19] combined concepts of vagueness expressed by membership degrees in fuzzy sets [117] and indiscernibility in RST to obtain fuzzy rough set theory (FRST). FRST allows partial membership of an object to the lower and upper approximations, and moreover, approximate equality between objects can be modeled by means of fuzzy indiscernibility relations. An advantage of this is that we do not need to perform discretization if our data contain real-valued attributes. FRST has been used e.g., for feature selection, instance selection, classification, and regression. There are many application areas that have been addressed by FRST, see e.g., [17,38,116].

[*] Corresponding author.
E-mail addresses: lala.s.riza@decsai.ugr.es (L.S. Riza), J.M.Benitez@decsai.ugr.es (J.M. Benítez).

http://dx.doi.org/10.1016/j.ins.2014.07.029
0020-0255/© 2014 Elsevier Inc. All rights reserved.

Software packages are essential for an effective deployment of these techniques as well as to facilitate further research. To date, several software packages for both theories are already available. Rough Set Data Explorer (ROSE) is an RST-based software system created by the Laboratory of Intelligent Decision Support Systems of the Institute of Computing Science in Poznań [81,82]. In [3,4], a free software system for data exploration, classification support, and knowledge discovery called the Rough Set Exploration System (RSES) was presented. The rough set toolkit for analysis of data (ROSETTA), which is an advanced system for RST data analysis [72,73], includes the RSES library as the computation kernel. Also, a few algorithms from FRST have been implemented in the Waikato Environment for Knowledge Analysis (WEKA) [42]. WEKA is a collection of machine learning algorithms for data mining tasks implemented in Java [33]. Rough Set Based Intelligent Data Analysis System (RIDAS) is another data mining toolkit utilizing notions from RST [109]. It was developed at Chongqing University of Posts and Telecommunications and consists of a C++ kernel and a GUI for Windows systems. An important data mining system for inducing decision rules from various types of data, called Learning from Examples based on Rough Sets (LERS), was created at University of Kansas [30]. There have also been developed a few rule based expert systems for a medical diagnostics purposes. One of the most prominent is PRIMEROSE [106] which allows generation of decision and inhibitory rules to construct reliable differential diagnosis.

Generally, the available software packages only consider the implementations of discretization, feature selection, and rule induction algorithms which can be used by practitioners to address their problems. However, there are no packages that provide comprehensive facilities for an implementation of fundamental concepts for academic purposes and further research. Therefore, this paper presents the *RoughSets* package that allows researchers and practitioners to explore both the basic knowledge of the theories and their applications. It was written mainly in the R language [39,104].

R is a widely used analysis environment for scientific computing and visualization, such as statistics, data mining, bioinformatics, and machine learning. Currently, over 5000 packages are included in the repositories of the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/ and the Bioconductor project at http://www.bioconductor.org/. Every package submitted into the repositories is checked to meet some quality standards, such as representative documentation and running on any operating systems (e.g., MS Windows, Mac OS X, Linux). Furthermore, according to a popularity survey conducted by Muenchen [66], R has been the most popular tool used for data analytics, data mining, and big data in 2013. In this setting, it seems quite natural to design and develop a complete and solid package for RST and FRST in R, which is the main motivation for the software that we present in this paper. The *RoughSets* package is available on CRAN at http://cran.r-project.org/package=RoughSets.

The remainder of this paper is structured as follows. Section 2 gives a brief overview of RST and FRST. Section 3 presents the main applications of both theories. In Section 4, we discuss the package architecture and capabilities of the package in detail. Section 5 shows examples of the usage of the package. We provide a review and comparison with currently available packages based on their capabilities and functionalities in Section 6. Finally, Section 7 concludes the paper.

2. Rough set and fuzzy rough set preliminaries

In this section, we review some basic notions related to RST and FRST. In particular, we focus on the indiscernibility relation, the lower and upper approximations, the positive and boundary region, and the decision-relative discernibility matrix. For further details on basic concepts and extensions of RST, interested readers are referred to [77–79], whereas some good introductions to FRST include e.g., [13,19,20,114,115].

We first introduce some notations which are used throughout the paper. A dataset is represented in terms of an information system $\mathcal{A} = (U, A)$ [74], where U is a finite, non-empty set of objects called the universe of discourse[1] and A is a finite, non-empty set of attributes, such that $a : U \to V_a$ for every $a \in A$, where V_a is the set of values that the attribute a may take. A decision system is a special kind of information system, used in the context of classification and prediction, in which d is a designated attribute called the decision attribute, and the attributes in A are called conditional attributes. More formally, it is a pair $\mathcal{A} = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. For RST and some of the FRST methods, the decision d has to be nominal. In the other FRST methods, the decision can be nominal or real-valued. The methods in the *RoughSets* package are implemented accordingly, allowing real-valued decisions where it is possible.

2.1. Indiscernibility relation

The main notion of RST is the indiscernibility relation.[2] Basically, it can be understood as a relation showing to what extent two objects are identical or similar. In the following subsections, we explain how to construct such relations in RST and FRST.

2.1.1. RST

Pawlak [75] considered an equivalence relation to model indiscernibility. Given an information system (U, A), for any $B \subseteq A$ the equivalence relation R_B is defined by

[1] In table format, it refers to all instances, experiments or rows.

[2] There exist also alternative interpretations of RST, where the notion of indiscernibility is replaced e.g. by the dominance relation (see e.g. [26]). These interpretations are not considered in this paper.