Information Sciences 000 (2016) 1–20

Contents lists available at ScienceDirect

Information Sciences

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

Feature selection and approximate reasoning of large-scale set-valued decision tables based on α-dominance-based quantitative rough sets

Hong-Ying Zhang*, Shu-Yun Yang

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaan'xi 710049, P.R. China

A R T I C L E I N F O

Article history:
Received 31 August 2015
Revised 16 June 2016
Accepted 19 June 2016
Available online xxx

Keywords:
Set-valued decision tables
α-dominance relation
Quantitative rough sets
Feature selection
Approximate reasoning
Disjunctive
Conjunctive

A B S T R A C T

Set-valued data are a common type of data for characterizing uncertain and missing information. Traditional dominance-based rough sets can not efficiently deal with large-scale set-valued decision tables and usually neglect the disjunctive semantics of sets. In this paper, we propose a general framework of feature selection and approximate reasoning for large-scale set-valued information tables by integrating quantitative rough sets and dominance-based rough sets. Firstly, we define two new partial orders for set-valued data via the conjunctive and disjunctive semantics of a set. Secondly, based on α-disjunctive dominance relation and α-conjunctive dominance relation defined by the inclusion measure, we present α-dominance-based quantitative rough set models for these two types of set-valued decision tables. Furthermore, we study the issue of feature selection in set-valued decision tables by employing α-dominance-based quantitative rough set models and discuss the relationships between the relative reductions and discernibility matrices. We also present approximate reasoning models based on α-dominance-based quantitative rough sets. Finally, the application of the approach is illustrated by some real-world data sets.

1. Introduction

The classical rough set theory, proposed by Pawlak [37,38], defines a pair of lower and upper approximations or equivalently three pair-wise disjoint positive, negative and boundary regions of a given set by using the set-inclusion relation and the nonempty set-overlapping condition. Rough sets evaluate the significance of attributes and derive the decision rules. The approximation operators of a Pawlak rough set ensure that both positive and negative regions are error-free. That is, no acceptance and rejection errors are allowed during the process of decision making. While we should allow some degrees of errors in dealing with large-scale data sets for the sake of flexible and practicability. Several generalized quantitative rough set models, which aim at making an acceptable decision for large-scale data sets, have been proposed based on an inclusion measures. They may be broadly classified into probabilistic and non-probabilistic approaches.

Probabilistic rough sets [39,53,54,60–65,76] quantify the set inclusion relation by a conditional probability, where the rough set approximations are presented by thresholds on the probability. Pawlak et al. [39] and Wong and Ziarko [53,54] first proposed probabilistic rough sets and introduced several models such as a 0.5-model [39,53] and a 0.5-β-model [55]. To

* Corresponding author. Tel.: +86 029 82663003.
E-mail addresses: zhyemily@mail.xjtu.edu.cn (H.-Y. Zhang), yangshuyun2009125@163.com (S.-Y. Yang).

http://dx.doi.org/10.1016/j.ins.2016.06.028
0020-0255/

overcome the lack of the interpretation and computation of the thresholds, Yao et al. [64,65] presented decision-theoretic rough sets for interpreting and computing a pair of thresholds based on a Bayesian decision theory. Yao [62] also proposed an approach to derive three-way decision rules based on decision-theoretic rough sets. The probabilistic rough sets are more general and flexible, which can be used to treat the objects from much larger universe and more insensitive to noises [63].

Non-probabilistic rough set models [9,28,40,48,58,66,71,72,74] are presented by considering other kinds of inclusion measures. Zhang and Leung [72,74] introduced the notion of inclusion degree to quantify the partial orders, which had been used to study non-probabilistic rough sets [58,71]. Xu et al. [58] investigated that many measures in rough sets, including accuracy measures and measures of dependence of attributes, are inclusion measures. Polkowski and Skowron [40,48] presented rough inclusion as a measure of subsethood relation. Gomolińska [9] gave a systematic research on rough-inclusion functions and their relationships with similarity measures and distance metrics. But the aforementioned studies about probabilistic and non-probabilistic approaches have not been unified, meanwhile insufficient attention has been paid to quantitative rough sets of non-probabilistic approaches. To fill this gap, Yao and Deng [66] initiated a study of quantitative rough set model which defines three regions by using inclusion measures with two thresholds and encompasses both probabilistic and non-probabilistic rough sets.

Set-valued information tables (SITs) [2,4,13,26,29,30,42,43], which can well characterize uncertainty information and missing information by using set-valued attributes. An incomplete information table can be regarded as a special kind of SITs [5,12] in which an object with a missing attribute value which can be replaced with the set of all possible values for this attribute. SITs have been analyzed by dominance-based rough sets [6,10,11,17,18,24,50,51,56,59,75] and fuzzy rough sets[4,5,8,16]. For instance, Shao and Zhang [47] presented a dominance-based rough set approach to reason in incomplete ordered information tables. Combining with fuzzy set theory [67], Dai and Tian [5] defined a fuzzy relation and constructed a fuzzy rough set model for SITs. Considering the ordered relations in SITs, Qian et al. [43,44] presented a rough set approach in set-valued ordered information tables (SOITs). Fan et al. [8] presented a dominance-based fuzzy rough set model for the decision analysis of an ordered uncertain and possibilistic data table. Greco et al. [10,11] presented a new fuzzy rough set approach based on the ordinal properties of fuzzy membership degrees. Incremental approaches of updating approximations and fast algorithms for computing approximations in set-valued approximations were proposed in [26,29,30]. Moreover, Inuiguchi et al. [17] introduced a variable-precision dominance-based rough set approach and studied the attribute reduction. Zhang et al. [70] proposed a general framework for the study of interval-valued decision tables by integrating the variable-precision-dominance-based rough set theory and the inclusion measure theory. However, disjunctive and conjunctive semantics of set-valued data have not been fully explored and the relative quantitative rough set models on SITs have not been studied. In this paper, on the basis of two new partial orders, we propose an α-dominance quantitative rough set model of SITs based on inclusion measures.

The rest of this paper is organized as follows. In Section 2, we mainly review the related work about inclusion measures and quantitative rough set model. In Section 3, we present two partial orders based on disjunctive and conjunctive semantics of a set. Furthermore, some inclusion measures on the partially ordered sets are constructed and a ranking model is proposed. In Section 4, we formulate an α-dominance-based quantitative rough set approach based on the inclusion measure and further study its properties. In Section 5, we investigate a framework of knowledge acquisition for SITs on the basis of α-dominance-based quantitative rough set model. Attribute reduction and the relationship between the reductions and the discernibility matrices are also analyzed. In Section 6, we further discuss approximate reasoning based on quantitative rough sets. In Section 7, we show some numerical experiments to demonstrate the application of the proposed approach. The paper is then ended with conclusions.

2. Preliminaries

In this section, we review some concepts of Pawlak's rough sets, dominance-based rough sets, inclusion measures and quantitative rough sets. One can refer to [10,37,38,46,66] for details.

2.1. Pawlak's rough sets

The theory of Pawlak rough sets [37,38] deals with inconsistent problems by separating of certain and doubtful knowledge. Now we present a slightly different formulation of Pawlak lower and upper approximations and the positive, negative and boundary regions by using the set-inclusion relation [62,63,66].

Let U be a finite nonempty set of objects and $R \subseteq U \times U$ be an equivalence relation on U, that is, R is reflexive, symmetric and transitive. The equivalence class containing x is defined as $[x]_R = \{y \in U | xEy\}$. The family of all the equivalence classes of R denoted by U/R constitutes a partition of U. An information table (IT) is a quadruple $S = (U, At, V, F)$ where U is a finite nonempty set of objects and At is a finite nonempty set of attributes. In S, the relation functions between the set of objects and the set of attributes are defined as f_a: $U \to V_a$ for any $a \in At$ where V_a is called the domain of an attribute a.

Definition 2.1. Let $S = (U, At, V, F)$ be an IT and $X \subseteq U$. The lower and upper approximations of X are defined by

$$\underline{R}(X) = \{x \in U | [x] \subseteq X\},$$
$$\overline{R}(X) = \{x \in U | \neg([x] \subseteq X^c)\}$$

Please cite this article as: H.-Y. Zhang, S.-Y. Yang, Feature selection and approximate reasoning of large-scale set-valued decision tables based on α-dominance-based quantitative rough sets, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.06.028