Knowledge-Based Systems 27 (2012) 443–450

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Uncertainty measurement for interval-valued decision systems based on extended conditional entropy

Jianhua Dai [a,b,*], Wentao Wang [a], Qing Xu [a], Haowei Tian [a]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] Center for the Study of Language and Cognition, Zhejiang University, Hangzhou 310028, China

ARTICLE INFO

Article history:
Received 18 May 2011
Received in revised form 24 August 2011
Accepted 23 October 2011
Available online 2 November 2011

Keywords:
Uncertainty measurement
Interval-valued decision systems
Rough set theory
Similarity degree
Conditional entropy
Rough decision entropy

ABSTRACT

Uncertainty measures can supply new points of view for analyzing data and help us to disclose the substantive characteristics of data sets. Some uncertainty measures for single-valued information systems or single-valued decision systems have been developed. However, there are few studies on the uncertainty measurement for interval-valued information systems or interval-valued decision systems. This paper addresses the uncertainty measurement problem in interval-valued decision systems. An extended conditional entropy is proposed in interval-valued decision systems based on possible degree between interval values. Consequently, a concept called rough decision entropy is introduced to evaluate the uncertainty of an interval-valued decision system. Besides, the original approximation accuracy measure proposed by Pawlak is extended to deal with interval-valued decision systems and the concept of interval approximation roughness is presented. Experimental results demonstrate that the rough decision entropy measure and the interval approximation roughness measure are effective and valid for evaluating the uncertainty measurement of interval-valued decision systems. Experimental results also indicate that the rough decision entropy measure outperforms the interval approximation roughness measure.

1. Introduction

Rough set theory, originally proposed by Pawlak and discussed in greater detail in [1,2], has become a popular approach for the joint management of uncertainty and vagueness and has been applied in many fields [3–7].

Pawlak [2] proposed two numerical measures *accuracy* and *roughness* to evaluate uncertainty of a rough set in information systems, as well as *approximation accuracy* of a rough classification in decision systems. Some efforts were attracted to extend the Pawlak's uncertainty model. Based on granulation, a measurement of uncertainty of a set in an information system and approximation accuracy of a rough classification in a decision table was proposed in [8].

Information entropy, proposed by Shannon [9] in information theory, has been an effective and powerful mechanism for characterizing the information content in diverse models. The measurement of uncertain information by entropy has been deployed in a wide range of fields. The extension of entropy and its variants were adapted for rough set in [10–21]. For example, Düntsch and Gediga defined the information entropy and three kinds of conditional entropies in rough sets for predicting a decision attribute [12]. Beaubouef et al. [13] proposed a method measuring uncertainty of rough sets and rough relation databases based on rough entropy. Wierman [11] presented the measures of uncertainty and granularity in rough set theory, along with an axiomatic derivation. Yao et al. [14] studied several kinds of information-theoretical measures for attribute importance in rough set theory. Liang et al. [16] proposed a new method for evaluating both uncertainty and fuzziness. Qian and Liang [19] proposed a combination entropy for evaluating uncertainty of a knowledge from an information system. However, the methods mentioned above are based on single-valued information systems.

Interval-valued information systems (or Interval information systems) are an important type of data tables, and generalized models of single-valued information systems [22]. Several authors have studied about interval-valued information systems and interval-valued decision systems [22–26]. Yao et al. [23,24] presented a model for the interval set by using the lower and upper approximations in interval-valued information systems, as well as introduced the generalized decision logic. Leung et al. [26] investigated a rough set approach to discover classification rules through a process of knowledge induction which selects decision rules with a minimal set of features in interval-valued information systems. Qian et al. [22] proposed a dominance relation to interval information systems. Yang et al. [25] presented a dominance rela-

* Corresponding author at: College of Computer Science, Zhejiang University, Hangzhou 310027, China.
E-mail address: jhdai@zju.edu.cn (J. Dai).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.10.013

tion and generated the optimal decision rules in incomplete interval-valued information system. Wu and Liu [27] introduced the real formal concept analysis about grey-rough set theory by using grey numbers, and proposed a grey-rough set approach to Galois lattices reductions. So far, however, there are few studies on the uncertainty measurement issue for interval-valued information systems (corresponding to unsupervised learning) or interval-valued decision systems (corresponding to supervised learning). In this paper, we address the uncertainty measurement issue in interval-valued decision systems and intend to construct effective uncertainty measures for interval-valued decision system. A similarity relation based on possible degree between two interval numbers is given, under which the concept of extended conditional entropy is proposed. Based on the proposed concept of conditional entropy, a measure of uncertainty for interval-valued decision systems called rough decision entropy is presented. Besides, the original approximation accuracy measure proposed by Pawlak is extended to deal with interval-valued decision systems and the concept of interval approximation roughness is presented. Experimental results demonstrate that the rough decision entropy measure and the interval approximation roughness measure are effective and valid for evaluating the uncertainty measurement of interval-valued decision systems. Experimental results also indicate that the rough decision entropy measure outperforms the interval approximation roughness measure.

This rest of the paper is organized as follows. Some basic concepts and notations in rough set theory are introduced in Section 2. In Section 3, serval key concepts of our method in interval-valued decision systems are illustrated in detail, including similarity degree, θ-conditional entropy, as well as θ-rough decision entropy. Some illustrative examples are also given. Simulation experiments are conducted to test and verify the effectiveness of the proposed measure in Section 4. Section 5 concludes the paper.

2. Preliminary knowledge

At first, some basic concepts in rough set theory are reviewed, including decision system, indiscernibility relation and approximation regions.

2.1. Indiscernibility relation and approximation regions

A decision system is a pair $\delta = (U, A \cup \{d\})$, where U is a non-empty finite set of objects called the universe of discourse; A is a non-empty finite set of attributes called condition attributes and d is class label called decision attribute; for any $a_\kappa \in A$, there exits a mapping $U \to V_{a_\kappa}$, where V_{a_κ} is called the value domain of a_κ.

For an attribute subset $B \subseteq A$,B determines a binary indiscernible relation denoted by $IND(B)$ as follows

$$IND(B) = \{(u_i, u_j) \in U^2 | \forall a_\kappa \in B, a_\kappa(u_i) = a_\kappa(u_j)\}$$

Note that the relation $IND(B)$ constitutes a partition of U, which is denoted by $U/IND(B)$ or U/B for simplification. In fact, U/B represents equivalence classes which are indiscernible by the attribute subset B.

For any given decision system $\delta = (U, A \cup \{d\})$, $B \subseteq A$ and $X \subseteq U$, one can define a lower approximation of X and an upper approximation of X in terms of U by

$$\underline{B}X = \{x \in U | [x]_{IND(B)} \subseteq X\}$$
$$\overline{B}X = \{x \in U | [x]_{IND(B)} \cap X \neq \emptyset\}$$

where $\underline{B}X$ is a set of objects that belong to X with certainty, while $\overline{B}X$ is a set of objects that possibly belong to X. X is called B-definable if and only if $\underline{B}X = \overline{B}X$. Otherwise, X is rough with respect to B when $\underline{B}X \neq \overline{B}X$.

Given the upper and lower approximations $\overline{B}X$ and $\underline{B}X$ of X, a subset of U, the B-positive region of X is $POS_B(X) = \underline{B}X$, the B-negative region of X is $NEG_B(X) = U - \overline{B}X$, and the B-boundary or B-borderline region of X is $NR_B(X) = \overline{B}X - \underline{B}X$. It is worth pointing out that since the decision attribute d is not taken into consideration, the upper and lower approximations are also applied to information systems without decision attributes.

2.2. Uncertainty measures in decision systems

Pawlak [2] proposed two numerical measures for evaluating uncertainty of a rough set X: *accuracy* and *roughness* in an information system. They are both modeled from the approximation regions, where elements of the upper approximation region have uncertain participation in the rough set and those in the lower approximation region have completely participation in the rough set.

However, *accuracy* and *roughness* can only be applied to information systems, since decision attribute is not taken into consideration. Therefore, *approximation accuracy* of a rough classification was introduced [2]. Let $U/d = \{D_1, D_2, \ldots, D_m\}$ be partitions constituted by decision attribute d on the universe U and the attribute subset $B \subseteq A$.

The approximation accuracy of U/d by B is defined as

$$\alpha_B(U/d) = \frac{\sum_{D_i \in U/d} |\underline{B}D_i|}{\sum_{D_i \in U/d} |\overline{B}D_i|}$$

where $|\cdot|$ denotes the number of elements of the set.

The approximation accuracy provides the percentage of possible correct decisions when discerning and classifying objects under the attribute subset B. Although, $\alpha_B(U/d)$ takes into account the number of elements in each of the approximation regions and evaluates well the uncertainty arising from the boundary region in some situations. However, there exist some limitations as pointed out in [8,13].

3. Uncertainty measurement in interval-valued decision systems

3.1. Similarity relation between two intervals

Ranking interval values are quite different from ranking real values [28–30]. In this section, a similarity measure based on possible degree is constructed to estimate two interval values.

Definition 1 [28]. Let $A = [a^-, a^+]$ and $B = [b^-, b^+]$ be two interval values. The possible degree of interval valued A relative to interval value B is defined as:

$$P_{(A \geqslant B)} = \min\left\{1, \max\left\{\frac{a^+ - b^-}{(a^+ - a^-) + (b^+ - b^-)}, 0\right\}\right\} \tag{1}$$

$P_{(A \geqslant B)}$ can be viewed as the possible degree of interval valued A greater than interval value B.

It is worth noting that $P_{(A \geqslant B)} \neq P_{(B \geqslant A)}$ in general.

Based on the concept of possible degree, we now define the similarity degree between two interval values.

Definition 2. Let $A = [a^-, a^+]$ and $B = [b^-, b^+]$ be two interval values. The similarity of the two interval values are defined as:

$$S_{AB} = 1 - |P_{(A \geqslant B)} - P_{(B \geqslant A)}| \tag{2}$$

where $P_{(A \geqslant B)}$ and $P_{(B \geqslant A)}$ are the possible degree of A relative to B and the possible degree of B relative to A respectively.