Knowledge-Based Systems 000 (2017) 1–13

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Uncertainty measurement for incomplete interval-valued information systems based on α-weak similarity

Jianhua Dai[a,b,*], Bingjie Wei[b], Xiaohong Zhang[c], Qilai Zhang[b]

[a] Key Laboratory of High Performance Computing and Stochastic Information Processing (Ministry of Education of China), College of Mathematics and Computer Science, Hunan Normal University, Changsha, Hunan 410081, China
[b] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[c] School of Arts and Sciences, Shaanxi University of Science & Technology, Xi'an 710021, China

ARTICLE INFO

Article history:
Received 31 December 2016
Revised 29 August 2017
Accepted 1 September 2017
Available online xxx

Keywords:
Incomplete interval-valued information
Rough sets
Uncertainty measure
Weak similarity

ABSTRACT

Rough set theory is a powerful mathematical tool to deal with uncertainty in data analysis. Interval-valued information systems are generalized models of single-valued information systems. Recently, uncertainty measures for complete interval-valued information systems or complete interval-valued decision systems have been developed. However, there are few studies on uncertainty measurements for incomplete interval-valued information systems. This paper aims to investigate the uncertainty measures in incomplete interval-valued information systems based on an α-weak similarity. Firstly, the maximum and the minimum similarity degrees are defined when interval-values information systems are incomplete based on the similarity relation. The concept of α-weak similarity relation is also defined. Secondly, the rough set model is constructed. Based on this model, accuracy, roughness and approximation accuracy are given to evaluate the uncertainty in incomplete interval-valued information systems. Furthermore, experimental analysis shows the effectiveness of the constructed uncertainty measures for incomplete interval-valued information systems.

1. Introduction

The uncertainty and vagueness of data play an important role in practical applications, so how to capture the uncertainty of information becomes more and more popular. Rough set theory, proposed by Pawlak [1,2], is a powerful mathematical tool to deal with uncertainty, granularity and incompleteness of knowledge in information systems. However, for the reason that classical rough set can only deal with complete and symbolic datasets, many scholars have developed the theory from different perspectives. With the swarming of extended models, they are popular in a wide range of fields [3–22], such as machine learning [23–29], pattern recognition [30–32], medical diagnosis [5] and decision making [33–36]. Beyond all doubt, rough set theory has become an efficient tool dealing with uncertainty and vagueness.

Among so many complex data, interval-valued data have attracted much attention of scholars from all over the world. Due to its rich semantic explanations and flexibility, interval-valued data have been widely used in economics analysis [37], machine learning [38], manufacturing processes [39] and so on. With the development of rough set theory, many results have been achieved to analysis the uncertainty of interval-value from different viewpoints. Dai et al. [40] proposed a θ-rough degree to measure the uncertainty of interval-value information systems. Furthermore, Dai et al. [41] addressed the extended conditional entropy to measure the uncertainty of interval-value decision systems which provided a new approach for decision rule evaluation and knowledge discovery. It is common to obtain missing values in real life for certain reasons, such as omissions, measurement malfunction and missing in storage. Roughly speaking, there are three main strategies for dealing with incomplete information systems [42]: (1) completion, i.e. the certain value takes the place of the unknown value, which could be the most common value, or the mean, or the median of all known values of the attribute. Yang et al. [43] presented a dominance relation and generated the optimal decision rules in incomplete interval-valued information systems. In this model, they chose the left smallest and right biggest value as lower bound and upper bound of the incomplete interval value based on certain attribute respectively. (2) deletion, i.e. we discard all the instances containing any unknown attribute value. Maybe we drop some important information in this way; (3) "best left alone", i.e. we consider the unknown values as a special symbol. For the

* Corresponding author at: School of Computer Science and Technology, Tianjin University, Tianjin 300350, China.
E-mail address: david.joshua@qq.com (J. Dai).

http://dx.doi.org/10.1016/j.knosys.2017.09.009
0950-7051/

first two strategies, the original structure of data is destroyed and the incomplete data couldn't be used sufficiently, hence, the third is more significative. Kryszkiewicz proposed a kind of tolerance relation [44,45] for incomplete information systems, however, it is only for single-valued decision systems rather than incomplete interval-valued information systems. This paper aims to construct uncertainty measures for incomplete interval-valued information systems by proposing an alpha-weak similarity relation.

The remainder of this paper is organized as follows. In Section 2, some basic knowledge of rough set theory are reviewed. In Section 3, we construct a α-weak similarity relation in incomplete interval-valued information systems and give the properties. In Section 4, we construct the extended rough set model based on α-weak similarity and study the properties of uncertainty measure. In Section 5, accuracy and approximation accuracy are tested in some real datasets. Section 6 concludes the whole paper.

2. Preliminary knowledge

In this section, we firstly review some basic concepts in rough set theory.

2.1. Basic concepts in rough set theory

For the information system $\delta = (U, A)$, where U denotes a non-empty finite set of objects, which is called the universe; A denotes a non-empty finite set of conditional attributes. Each attribute subset $B \subseteq A$ determines a binary indiscernible relation as follows:

$$IND(B) = \left\{\left(u_i, u_j\right) \in U^2 | \forall a \in B, a(u_i) = a\left(u_j\right)\right\} \tag{1}$$

By the relation $IND(B)$, we obtain the partition of U denoted by $U/IND(B)$ or U/B. For $B \subseteq A$ and $X \subseteq U$, the lower approximation and the upper approximation of X can be defined as follows [1]:

$$\underline{B}X = \{u_i \in U | [u_i] \subseteq X\}$$

$$\overline{B}X = \{u_i \in U | [u_i] \cap X \neq \emptyset\}$$

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. If $\underline{B}X = \overline{B}X$, X is named B-definable. Otherwise, X is named B-rough. Based on $\underline{B}X$ and $\overline{B}X$, the B-positive region, B-negative region and B-borderline region of X are defined respectively as follows:

$$POS_B(X) = \underline{B}X$$

$$NEG_B(X) = U - \overline{B}X$$

$$BN_B = \overline{B}X - \underline{B}X$$

Pawlak proposed two numerical measures for evaluating the uncertainty of the given object set X: accuracy and roughness, which can be denoted as follows:

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|}$$

$$\rho_B(X) = 1 - \alpha_B(X) = 1 - \frac{|\underline{B}X|}{|\overline{B}X|}$$

where $|\cdot|$ denotes the number of collection elements. The accuracy and roughness describe the completeness and incompleteness in the knowledge about the given object set X respectively.

2.2. Incomplete interval-valued information systems and incomplete interval-valued decision systems

An interval-valued information system is a quadruple $IIS = \langle U, A, V, f\rangle$, where U is a nonempty set with a finite number of objects, called the universe; A is a nonempty finite set of conditional attributes; V is the set of interval-value domain of attributes; f is an information function which allocates values by domains of attributes to objects, for instance, $\forall a \in A,\ x \in U,\ f(a, x) \in V$, where $f(a, x)$ denotes the value of attribute a of object x. If there exists an $a \in A,\ x \in U$ such that $f(a, x)$ is an incomplete interval value containing missing bound(s) (missing the lower bound ([*, x]), missing the upper bound ([x, *]) or missing both the lower and upper bounds ([*, *])), then the interval-valued information system is called an Incomplete Interval-valued Information System (IIIS). Otherwise, it is called a complete interval-valued information system. Thus, an IIIS can be denoted as: $IIIS = \langle U, A, V, f\rangle$, where [*, x], [x, *], [*, *] $\in V$.

A decision system is a quadruple $IIDS = \langle U, C \cup \{d\}, V, f\rangle$, where d is the decision attribute set, C is the conditional attribute set. If there exists an $a \in A,\ x \in U$ such that $f(a, x)$ is an interval value containing missing bound(s) (missing the lower bound ([*, x]), missing the upper bound ([x, *]) or missing both the lower and upper bounds([*, *])), then the decision system is called an Incomplete Interval-valued Decision System (IIDS). Thus, an IIDS can be denoted as: $IIDS = \langle U, C \cup \{d\}, V, f\rangle$, where [*, x], [x, *], [*, *] $\in V$.

2.3. Tolerance relation in IIS

Definition 1 [44,45]**.** Given an incomplete information system $IIS = \langle U, A, V, f\rangle, * \in V = \cup_{a \in A} V_a$, for any subset of attributes $B \in A$, let $T^{IIS}(B)$ denote the binary tolerance relation between objects that are possibly indiscernible in terms of values of attribute B. $T^{IIS}(B)$ is defined as

$$\begin{aligned} T^{IIS}(B) = \{(x, y) | \forall a \in B, f(a, x) \\ = f(a, y) \vee f(a, x) = * \vee f(a, y) = *\} \end{aligned} \tag{2}$$

$T^{IIS}(B)$ is reflexive and symmetric, but not transitive.

Definition 2 [46]**.** The tolerance class of an object x with respect to an attribute set B is defined by:

$$\begin{aligned} T^{IIS}{}_B(x) = \{y | (x, y) \in T^{IIS}(B)\} = \{y | \forall a \in B, f(a, x) \\ = f(a, y) \vee f(a, x) = * \vee f(a, y) = *\} \end{aligned} \tag{3}$$

3. α-weak similarity relation for incomplete interval valued information systems

Unlike real values, it is difficult to compare two interval values using traditional methods. Enlightened by the similarity measure for general interval-valued data proposed in [47], we give a definition of similarity between two intervals.

Definition 3. Let $U = \{u_1, u_2, \ldots, u_n\}$ be the universe of interval values, $\forall u_i,\ u_j \in U,\ u_i = [u_i^-, u_i^+]$ and $u_j = \left[u_j^-, u_j^+\right],\ u_i^- < u_i^+$ or $u_j^- < u_j^+$. The similarity degree of the interval value u_i relative to the interval value u_j is defined as follows:

$$S_{ij} = 1 - \frac{1}{2} * \frac{|u_i^+ - u_j^+| + |u_i^- - u_j^-|}{\max(u_i^+, u_j^+) - \min(u_i^-, u_j^-)} \tag{4}$$

As for the relation between intervals $[a^-, a^+]$ and $[b^-, b^+]$, there are six types of situations, shown in Fig. 1. Hence, according to Definition 3, we can compute the similarity degree S easily as follows:

(a) $a^- < a^+ \leq b^- < b^+$ and $S = 1 - \frac{1}{2} * \frac{(b^- - a^-) + (b^+ - a^+)}{b^+ - a^-}$

Please cite this article as: J. Dai et al., Uncertainty measurement for incomplete interval-valued information systems based on α-weak similarity, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.09.009