Information Sciences 251 (2013) 63–78

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Uncertainty measurement for interval-valued information systems

Jianhua Dai [a,b,*], Wentao Wang [a], Ju-Sheng Mi [b,c]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang 050024, China
[c] College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050016, China

ARTICLE INFO

Article history:
Received 1 May 2011
Received in revised form 7 June 2013
Accepted 16 June 2013
Available online 25 June 2013

Keywords:
Interval data
Uncertainty measure
Rough set theory
Interval-valued information systems
Similarity degree
Roughness

ABSTRACT

Interval-valued information systems are generalized models of single-valued information systems. Accuracy and roughness are employed to depict the uncertainty of a set under an attribute subset in a Pawlak rough set model based on equivalence classes. Information-theoretic measures of uncertainty for rough sets have also been proposed. However, there are few studies on uncertainty measurements for interval-valued information systems. This paper addresses the uncertainty measurement problem in interval-valued information systems. The concept of the similarity degree, based on the possible degree, is introduced. Consequently, the similarity relation between two interval objects are constructed by a given similarity rate θ. Based on the similarity relation, θ-similarity classes are defined. Under this definition, θ-accuracy and θ-roughness are given for interval-valued information systems, which are generalizations of the concepts *accuracy* and *roughness* for the equivalence relation-based rough set model. Moreover, an alternative uncertainty measure, called the θ-rough degree, is proposed. Theoretical studies and numerical experiments show that the proposed measures are effective and suitable for interval-valued information systems.

1. Introduction

Rough set theory, originally proposed by Pawlak [42–46], is an important approach for the joint management of uncertainty and vagueness. The theory has been applied successfully in many areas, such as data mining, data analysis and decision making [8,9,11,13,17,18,20,25,26,33,41,53,55,58,61,65].

In rough set theory, the accuracy measure and the rough measure are important numerical characterizations that quantify the imprecision of a rough set caused by its boundary region. Another of the most significant concepts in rough set theory is the uncertainty of a set. Pawlak [43] proposed two numerical measures, *accuracy* and *roughness* to evaluate the uncertainty of a rough set. Liang et al. [32] proposed a new measure of uncertainty that is based on knowledge granulation to improve the measures in [43]. Recently, Dai et al. [16] studied uncertainty measures in incomplete information systems by pure rough set approach.

Information theory, introduced by Shannon [52] for communication theory, has been a useful and powerful mechanism for characterizing information content in diverse models. It is believed by Shannon that the physical entropy used in thermodynamics is more or less closely related to the concept of information used in communication theory. Therefore, he defined the information entropy to provide a measure of uncertainty. The measurement of uncertain information by entropy

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

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.06.047

has been deployed in a wide range of fields. The extension of entropy and its variants was introduced into rough set theory in [1,12,14,15,21,30,31,47,56,63,64]. For example, Duntsch and Gediga first defined information entropy and three types of conditional entropies in rough sets for predicting a decision attribute [21]. Beaubouef et al. [1] proposed a method for measuring the uncertainty of rough sets and rough relation databases that is based on rough entropy. Wierman [56] presented measures of uncertainty and granularity in rough set theory, along with an axiomatic derivation. Yao [63] studied several types of information-theoretical measures for attribute importance in rough set theory. Liang et al. [30] proposed a new method for evaluating both uncertainty and fuzziness. Qian and Liang [47] proposed a combination entropy for evaluating uncertainty. Bianucci et al. [2,3] explored entropy and co-entropy approaches for uncertainty measurements of coverings. However, the methods mentioned above are based on single-valued information systems.

Intervals appear to a way to describe the uncertainty that affects the observed values. The uncertainty can be considered to be the incapability of obtaining true values under existing conditions. Uncertainty can be the expression of three causes: randomness, vagueness or imprecision [27]. Interval analysis was introduced by Moore in 1959 as a tool for the automatic control of the errors in a computed result [36], where an interval number is considered to be an extension of the real numbers. An interval signifies the extent of tolerance (or a region) that the parameter can possibly take. Interval analysis has been widely used, especially in solving Linear Programming problems with interval coefficients [39,40,51,66].

Lipski conducted some fundamental research on interval information systems [34,35]. His incomplete information databases are actually information systems with interval values. Interval data have also been studied in symbolic data analysis (SDA) [6], a new domain related to multivariate analysis, pattern recognition and artificial intelligence. In SDA, many research studies have been performed to extend classical data analysis and statistical methods to symbolic data [4,5]. Gowda and Diday proposed a symbolic clustering of interval data [23]. Souza et al. [54] proposed an adaptive dynamic clustering algorithm for interval data based on City-block distances. Irpino et al. [27] constructed dynamic clustering of interval data using a Wasserdtein-based distance. De Carvalho et al. [19] introduced partitional dynamic clustering methods based on single adaptive city-block and Hausdorf distances. Recently, Hedjazi et al. [24] have proposed a similarity measure for interval-valued data therein for interval feature selection.

To handle interval-valued data, researchers in rough set theory have studied interval-valued information systems as generalized models of single-valued information systems. Yao [62] presented a model for the interval set by using the lower and upper approximations in interval-valued information systems and, in additional, introduced the generalized decision logic. Dai [10] investigated the algebraic structures for interval-set-valued rough sets generated from an approximation space. Leung et al. [29] investigated a rough set approach to discover classification rules through a process of knowledge induction that selects decision rules with a minimal set of features in interval-valued information systems. Qian et al. [48] proposed a dominance relation for interval information systems. Yang et al. [60] presented a dominance relation and generated the optimal decision rules in incomplete interval-valued information systems. Yamaguchia et al. [59] defined a grey-rough set model for interval data in which the grey-rough approximation was based on the grey lattice relation instead of an equivalence relation. Wu et al. [57] introduced the real formal concept analysis of grey-rough set theory by using grey numbers, and proposed a grey-rough set approach to Galois lattice reductions. Actually, formal concept analysis alone can be generalized to address intervals [28]. Sakai et al. [50] developed a prototype system for rule generation in Lipski's incomplete information databases that can handle interval information systems. Recently, Dai et al. [15] constructed an extended conditional entropy for interval-valued decision systems.

So far, however, there are few studies on uncertainty measurements for interval-valued information systems. In this paper, we address the uncertainty measurement issue in interval-valued information systems. A similarity relation that is based on the possible degree between two interval numbers is given, under which the generalized measures of accuracy and roughness for interval-valued information systems are presented. Moreover, an alternative uncertainty measure is proposed. Theoretical studies and experimental results show that the proposed measures are effective and suitable for evaluating the uncertainty for interval-valued information systems.

The remainder of this paper is organized as follows. Some preliminaries of rough set theory are introduced in Section 2. In Section 3, the similarity degree between two interval values and the similarity relation based on possible degrees are constructed. In Section 4, extensions of the concepts accuracy and roughness for interval-valued information systems are given and studied. An alternative measure for uncertainty is also proposed and investigated. In Section 5, experiments on real world data sets and simulation data sets, to display the feasibility of the studied methods, are conducted. Section 6 concludes the paper.

2. Preliminaries

2.1. Indiscernibility relation and approximation regions

An information system is a pair $\delta = (U,A)$, 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; for any $a_\kappa \in A$, there exists a mapping $U \to V_{a_\kappa}$, where V_{a_κ} is called the valued set of a_κ.

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

$$IND(B) = \{(u_i, u_j) \in U^2 | \forall a_\kappa \in B, a_\kappa(u_i) = a_\kappa(u_j)\} \quad (1)$$