Knowledge-Based Systems 27 (2012) 126–136

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Distance: A more comprehensible perspective for measures in rough set theory

Jiye Liang, Ru Li, Yuhua Qian *

School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, Shanxi, China

A R T I C L E I N F O

Article history:
Received 6 October 2010
Received in revised form 26 September 2011
Accepted 3 November 2011
Available online 6 December 2011

Keywords:
Rough set theory
Set distance
Partition distance
Measures
Information granularity

A B S T R A C T

Distance provides a comprehensible perspective for characterizing the difference between two objects in a metric space. There are many measures which have been proposed and applied for various targets in rough set theory. In this study, through introducing set distance and partition distance to rough set theory, we investigate how to understand measures from rough set theory in the viewpoint of distance, which are inclusion degree, accuracy measure, rough measure, approximation quality, fuzziness measure, three decision evaluation criteria, information measure and information granularity. Moreover, a rough set framework based on the set distance is also a very interesting perspective for understanding rough set approximation. From the view of distance, these results look forward to providing a more comprehensible perspective for measures in rough set theory.

1. Introduction

Rough set theory proposed by Pawlak in [17] is a relatively new soft computing tool for the analysis of a vague description of an object, and has become a popular mathematical framework for pattern recognition, image processing, feature selection, neuro computing, conflict analysis, decision support, data mining and knowledge discovery from large data sets [1–3,13,23,30,39,42,44]. Rough-set-based data analysis starts from a data table, called information tables. The information tables contain data about objects of interest, characterized by a finite set of attributes. It is often interesting to discover some dependency relationships (patterns). An information table where condition attributes and decision attributes are distinguished is called a decision table. From a decision table one can induce some patterns in form of "*if*..., *then*..." decision rules [5,6,19,30]. More exactly, the decision rules say that if condition attributes have given values, then decision attributes have other given values.

To date, many measures for uncertainty have been proposed in rough set theory. As follows, for our further development, we briefly review several important measures. The concept of inclusion degree has been introduced into rough set theory, which is derived from the including measure among sets. Several authors have established several important relationships between inclusion degree and measures of rough set data analysis [27,40]. In rough set theory, as three classical measures, approximation accuracy, rough measure and approximate quality can be used to assess the roughness of a rough set and a rough classification [7,17]. For any object on a given universe, the membership function of the object in a rough set can be derived by the inclusion degree between the equivalence class including itself and a target concept, which can construct a fuzzy set on the universe. Several authors have studied the fuzziness of a rough set from various viewpoints [21,41]. In recent years, how to evaluate the decision performance of a decision rule and a decision-rule set has become a very important issue in rough set theory. There are two classical measures such as certainty measure and coverage measure [17]. In order to assess the decision performance of a decision table, Qian et al. [20] proposed three evaluation parameters α, β and γ which are used to calculate the entire certainty, the entire consistency and the entire support of all decision rules from a given decision table. However, each of the above measures is defined by different forms, which is hard to understand their sematic meanings. In other word, the uniform characterization of these measures is desirable. As we know, the concept of distance is a main approach to understand the difference between two objects in algebra, geometry, set theory, coding theory and many other areas. Hence, in this study, we aims to propose the concept of set distance to characterize and redefine each of these measures in order to more easily comprehend their meanings. It is exciting that Pawlak's rough set framework can be reconfigured using the set distance. This idea also can be used to redefined the variable precision rough set model proposed by Ziarko [46]. These results will be very helpful for us to understand the essence of rough set approximation. That is to say, it is a more comprehensible perspective for measures in rough set theory.

* Corresponding author at: School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China. Tel./fax: +86 0351 7018176.
E-mail addresses: ljy@sxu.edu.cn (J. Liang), liru@sxu.edu.cn (R. Li), jinchengqyh@126.com (Y. Qian).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.11.003

In addition, information entropy and information granularity are two main approaches to characterizing the uncertainty of an information system [14,18,28,43,45]. In recent years, several various forms of information entropy and information granularity have been given in [13,14,28,43,45]. It is deserved to point out that when the information granularity (or information entropy) of one equivalence partition is equal to that of the other equivalence partition, these two equivalence partitions have the same uncertainty. Nevertheless, it does not mean that these two equivalence partitions are equivalent. That is to say, information entropy and information granularity cannot characterize the difference between any two equivalence partitions in a given information table. In fact, we often need to distinguish two equivalence partitions for uncertain data processing in some practical applications. To date, how to measure the difference between equivalence partitions has not been reported. To further investigate uncertainty theory in the framework of rough set theory, for this consideration, we will propose the concept of partition distance to calculate the difference between two partitions on the same universe in this paper. In particular, we also reveal the essence of definitions of information entropy and information granularity from the viewpoint of partition distance.

The rest of this paper is organized as follows. Some preliminary concepts in rough set theory are briefly recalled in Section 2. In Section 3, we introduce the concept of set distance to characterize several important measures, which are inclusion degree, accuracy measure, rough measure, approximation quality, several decision evaluation parameters and the fuzziness measures of rough sets. In addition, we employ the set distance for reconfiguring the rough set framework and the variable precision rough set model. In Section 4, we first define the concept of partition distance to calculate the difference between two partitions on the same universe, then employ the partition distance to understand information entropy and information granularity from the viewpoint of distance. Section 6 concludes this paper with some remarks and discussions.

2. Preliminary knowledge in rough sets

In this section, we review some basic concepts such as indiscernibility relation, partition, partial relation of knowledge and decision tables in rough set theory.

An information table (sometimes called a data table, an attribute-value system, a knowledge representation system, etc.), as a basic concept in rough set theory, provides a convenient framework for the representation of objects in terms of their attribute values. An information table S is a pair (U,A), where U is a non-empty, finite set of objects and is called the universe and A is a non-empty, finite set of attributes. For each $a \in A$, a mapping $a: U \to V_a$ is determined by a given decision table, where V_a is the domain of a.

Each non-empty subset $B \subseteq A$ determines an indiscernibility relation in the following way,

$$R_B = \{(x,y) \in U \times U | a(x) = a(y), \forall a \in B\}.$$

The relation R_B partitions U into some equivalence classes given by

$$U/R_B = \{[x]_B | x \in U\}, \text{ just } U/B,$$

where $[x]_B$ denotes the equivalence class determined by x with respect to B, i.e.,

$$[x]_B = \{y \in U | (x,y) \in R_B\}.$$

Given an equivalence relation R on the universe U and a subset $X \subseteq U$. One can define a lower approximation of X and an upper approximation of X by

$$\underline{R}X = \{x \in U | [x]_R \subseteq X\}$$

and

$$\overline{R}X = \{x \in U | [x]_R \cap X \neq \emptyset\},$$

respectively [15]. The ordered pair $(\underline{R}X, \overline{R}X)$ is called a rough set of X with respect to R.

We define a partial relation $\preceq$ on the family $\{U/B | B \subseteq A\}$ as follows: $U/P \preceq U/Q$ (or $U/Q \succeq U/P$) if and only if, for every $P_i \in U/P$, there exists $Q_j \in U/Q$ such that $P_i \subseteq Q_j$, where $U/P = \{P_1, P_2, \ldots, P_m\}$ and $U/Q = \{Q_1, Q_2, \ldots, Q_n\}$ are partitions induced by P, $Q \subseteq A$, respectively. In this case, we say that Q is coarser than P, or P is finer than Q. If $U/P \preceq U/Q$ and $U/P \neq U/Q$, we say Q is strictly coarser than P (or P is strictly finer than Q), denoted by $U/P \prec U/Q$ (or $U/Q \succ U/P$).

It is clear that $U/P \prec U/Q$ if and only if, for every $X \in U/P$, there exists $Y \in U/Q$ such that $X \subseteq Y$, and there exist $X_0 \in U/P$, $Y_0 \in U/Q$ such that $X_0 \subset Y_0$.

A decision table is an information table $S = (U, C \cup D)$ with $C \cap D = \emptyset$, where an element of C is called a condition attribute, C is called a condition attribute set, an element of D is called a decision attribute, and D is called a decision attribute set. One can extract certain decision rules from a consistent decision table and uncertain decision rules from an inconsistent decision table.

3. Set distance and some measures in rough sets

The concept of set closeness between two classical sets is used to measure the degree of the sameness between sets. Let X and Y be two finite sets, the measure is defined by $H(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$ $(X \cup Y \neq \emptyset)$. Obviously, the formula $1 - H(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$ can characterize the difference between two finite classical sets. In a broad sense, this measure can be regarded as a generalized distance [22]. Using the measure, one can obtain the following distance between two finite classical sets.

Definition 1. Let X, Y are two finite sets. The distance between X and Y is defined as

$$d(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}, \quad (1)$$

where $X \cup Y \neq \emptyset$.

From the definition of the distance, one can easily obtain the following property.

Property 1. *The distance d satisfies the following properties:*

(1) $d(X,Y) \geqslant 0$;
(2) $d(X,Y) = d(Y,X)$;
(3) $d(X,Y) + d(Y,Z) \geqslant d(X,Z)$.

Proof. The three properties will be proved as follows.

(1) Obviously, $|X \cup Y| \geqslant |X \cap Y|$. Thus we have that,

$$d(X,\ Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} \geqslant 0.$$

(2) It is easy to know that $|X \cup Y| = |Y \cup X|$ and $|X \cap Y| = |Y \cap X|$. Therefore,

$$d(X,\ Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} = \frac{|Y \cup X| - |Y \cap X|}{|Y \cup X|} = d(Y,\ X).$$

(3) Given any a, b and c, and let $0 < b \leqslant a$, $c \geqslant 0$. From $\frac{b+c}{a+c} - \frac{b}{a} = \frac{c(a-b)}{a(a+c)} \geqslant 0$, it follows that $\frac{b}{a} \leqslant \frac{b+c}{a+c}$. Hence,