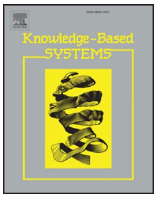
Knowledge-Based Systems 000 (2017) 1–10

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Measures of uncertainty for neighborhood rough sets

Yumin Chen[a,*], Yu Xue[b], Ying Ma[a], Feifei Xu[c]

[a] School of Computer & Information, Xiamen University of Technology, Xiamen 361024, China
[b] School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[c] School of Computer Science & Technology, Shanghai University of Electric Power, Shanghai 200090, China

ARTICLE INFO

Article history:
Received 15 May 2016
Revised 3 January 2017
Accepted 4 January 2017
Available online xxx

Keywords:
Rough sets
Neighborhood rough sets
Uncertainty measures
Neighborhood entropy
Information quantity

ABSTRACT

Uncertainty measures are critical evaluating tools in machine learning fields, which can measure the dependence and similarity between two feature subsets and can be used to judge the significance of features in classifying and clustering algorithms. In the classical rough sets, there are some uncertainty tools to measure a feature subset, including accuracy, roughness, information entropy, rough entropy, etc. These measures are applicable to discrete-valued information systems, but not suitable to real-valued data sets. In this paper, by introducing the neighborhood rough set model, each object is associated with a neighborhood subset, named a neighborhood granule. Several uncertainty measures of neighborhood granules are proposed, which are neighborhood accuracy, information quantity, neighborhood entropy and information granularity in the neighborhood systems. Furthermore, we prove that these uncertainty measures satisfy non-negativity, invariance and monotonicity. The maximum and minimum of these measures are also given. Theoretical analysis and experimental results show that information quantity, neighborhood entropy and information granularity measures are better than the neighborhood accuracy measure in the neighborhood systems.

1. Introduction

Uncertainty almost exists in everywhere of the real world, including randomness, fuzziness, vagueness, incompleteness and inconsistency. The question of representing and measuring the uncertain knowledge has become one of most important issue in the research of machine learning [10,46]. Good uncertainty measures applied to evaluating the information systems or decision support systems can improve the accuracies and efficiencies of the clustering and classification algorithms [36,50] in machine learning. Uncertainty measures have been widely used in many fields, such as machine learning [33,39], pattern recognition [1,11], information retrieval [34,37], image processing [29,38], medical diagnosis [12] and data mining [8].

Rough set theory [30], pioneered by the Polish scientist Pawlak Z. in 1982, is a valid paradigm in mathematics to tackle the imprecise, uncertain and tremendous data. Recently, many uncertainty measures established in rough sets require to tackle the widely existences of fuzzy data. Pawlak [31] employed the ratio of upper and lower approximations, named accuracy, to measure the uncertainty of a rough set in an information table. The approximation accuracy is used by Pawlak to evaluate the uncertainty of a rough classification in a decision system. The accuracy and approximation accuracy are monotone increasing functions with the growing of an attribute set. Pawlak also put forward the roughness and approximate roughness for measuring a rough set or a rough classification, which are two monotone decreasing functions. However, the Pawlak's uncertainty measures are not meticulous, since there is a case that the two different equivalence class sets probably have the same accuracy or roughness. Therefore, many scholars have proposed many measures to overcome the weakness from different perspectives, such as information quality [17], approximate quality [6], knowledge granularity [27], information entropy [26], combination of accuracy and granularity [43], etc. Miao and Hou [28] and Liang et al. [19] introduced various entropy measures into the field of rough sets, which provide some more effective and meticulous measurement tools, including information entropy [19,28], combination entropy [32], mutual information [40] and rough entropy [19]. Yao demonstrated various measures from multi-granularity views [43,46] and discussed their applications to KDD fields [42]. As these measures mainly aim to evaluate equivalence relations and equivalence classes, other authors have presented uncertainty measures and applications for fuzzy rough sets [7,13,14], probabilistic rough sets [20,22,48], covering rough sets [3,52], cost-sensitive [24,25] and neighborhood rough sets [49,51].

[*] Corresponding author.
E-mail address: cym0620@163.com (Y. Chen).

http://dx.doi.org/10.1016/j.knosys.2017.01.008
0950-7051/© 2017 Elsevier B.V. All rights reserved.

These measures for rough sets have been widely used in the researches and applications of machine learning and data mining. The classical rough set model is mainly applicable to the decision system with discrete data. As for widely existing continuous data, a discretization should be implemented. However, this preprocessing will result in the loss of information, reducing the classification accuracy. The notion of neighborhood based rough set was firstly introduced by Yao [44] based on neighborhood defined by a binary relation. The use of a distance function to define a neighborhood was put forward by Yao [45]. The neighborhood rough set model was utilized to classification by Hu et al. [15], which can handle the knowledge classification system with not only continuous data but also discrete data. There is a vast literature on attribute reduction [21], feature selection and rule extracting [16], classification and cluster [41], gene selection [23], image processing [2,47] and other fields [4]. Since the neighborhood relation is not the same as an equivalence relation, the classical uncertainty measurement tools and methods are not applicable to the neighborhood knowledge classification system. Considering the characteristics of continuous data, we use the neighborhood rough set model to granulate those data. Moreover, we study the uncertainty measures in neighborhood rough sets. After introducing several measures in classical rough sets, we propose naturally extensional measures of uncertainty for a neighborhood system. Firstly, by granulating the neighborhood information system, we construct some neighborhood class sets. Secondly, we define the concepts of neighborhood accuracy and neighborhood roughness to evaluate the uncertainty of the neighborhood class sets in a neighborhood information system. And we develop the terminologies of approximate accuracy and approximate roughness to estimate the uncertainty of a neighborhood decision system. Furthermore, the information quantity-based, the neighborhood entropy-based, and the information granularity-based measures are proposed. The theoretical analysis and several experiments show that the three proposed measures in the neighborhood system are better than the accuracy-based measure.

The remainder of this paper is structured as follows. An introduction to rough sets and several measures are presented in Section 2. In Section 3, the neighborhood rough set model is introduced. Then we propose four different uncertainty measurement methods, which are accuracy-based, information quantity-based, neighborhood entropy-based and information granularity-based measures. Furthermore, we prove some theorems associated with the uncertainty measures, contributing to understand information entropy and information granularity in a neighborhood system. In Section 4, some experiments are implemented to indicate the effectiveness of our proposed measures. Finally, this paper is concluded with some remarks and discussions in Section 5.

2. Rough set theory and its uncertainty measures

In this section, we present some descriptions and terminologies for the rough set model, which mainly include information system, equivalence relation, lower and upper approximations. Furthermore, we also review some uncertainty measures of an information system which can be found in [31].

2.1. Concepts of rough set model

The concept of information system is structured as a formalization knowledge representation. Generally, an information system is a quadruple represented by $IS = (U, A, V, f)$, consisting of a nonempty finite object set U, a nonempty finite attribute set A, a value domain union V and a mapping function f. Among them, $V = \bigcup_{a\in A} V_a$ for V_a denotes the value domain of attribute a, and any $a \in A$ determines a function f_a: $U \to V_a$. In a particular situation with $A = C \cup \{d\}$, a decision system is formalized by $DS = (U, C \cup \{d\}, V, f)$, where C is a set of condition attributes, and d is a decision attribute.

For any subset $P\subseteq A$, there is an indiscernibility relation represented by $IND(P)$ in the following:

$$IND(P) = \{(x, y) \in U \times U|\forall p \in P, f(x, p) = f(y, p)\}.$$

The $IND(P)$ satisfies reflexivity, symmetry and transitivity. So, it is an equivalence relation. The $[x]_P$ is an equivalence class of an object x related to the equivalence relation $IND(P)$. The set of all equivalence classes of $IND(P)$ is denoted by $U/IND(P)$, or simply U/P.

Given an information system $IS = (U, A, V, f)$ and an attribute subset $P\subseteq A$, for an object subset $X\subseteq U$, the lower and upper approximations of X with respect to P are defined in the follows:

$$P_*(X) = \{x \in U|[x]_P \subseteq X\},$$
$$P^*(X) = \{x \in U|[x]_P \bigcap X \neq \emptyset\}.$$

The tuple $< P_*(X), P^*(X) >$ is called a rough set, if the lower approximation is not equal to the upper approximation. If they are equal, the rough set degrades into a crisp set. Given an information system $IS = (U, A, V, f)$, for two subsets P, $Q\subseteq A$, U/P and U/Q are two partitions of U generated by $IND(P)$ and $IND(Q)$ respectively, then the positive, negative and boundary regions are defined in the follows:

$$POS_P(Q) = \bigcup_{X\in U/Q} P_*(X),$$
$$NEG_P(Q) = U - \bigcup_{X\in U/Q} P^*(X),$$
$$BND_P(Q) = \bigcup_{X\in U/Q} P^*(X) - \bigcup_{X\in U/Q} P_*(X).$$

2.2. Uncertainty measures of rough set model

In rough set model, there are two kinds of uncertainty measure methods to evaluate a knowledge representation system, which are accuracy-based method and entropy-based method. As for accuracy-based measures, there are four numerical measures including accuracy, approximation accuracy, roughness and approximation roughness. As for entropy measures, there are mainly three measures which are information entropy, conditional entropy and mutual information.

The accuracy is the ratio of the lower and upper approximation to measure the imprecision of a rough set. The roughness is an inverse of accuracy by a subtraction in the following.

Definition 1 [31]**.** Suppose $IS = (U, A, V, f)$ be an information system, for any object subset $X\subseteq U$ and any attribute subset $P\subseteq A$, the accuracy and roughness of X related to P are defined as follows:

$$\alpha_P(X) = \frac{|P_*(X)|}{|P^*(X)|}, \tag{1}$$

$$\rho_P(X) = 1 - \alpha_P(X). \tag{2}$$

The measures of accuracy and roughness are used for evaluating a rough set of information systems. However, those measures are not suitable for a rough classification in a decision system. Therefore, approximation accuracy is proposed by Pawlak [31] for evaluating the rough classification in a decision system.

Definition 2 [31]**.** Suppose $DS = (U, C \cup \{d\}, V, f)$ be a decision system, and $U/d = \{D_1, D_2, \ldots, D_m\}$ be equivalence classes generated by a decision attribute d on the universe U. For a condition attribute subset $P\subseteq C$, the approximation accuracy and approximation