Knowledge-Based Systems 57 (2014) 28–40

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A method for extracting rules from spatial data based on rough fuzzy sets

Hexiang Bai [a,b,c], Yong Ge [b,*], Jinfeng Wang [b], Deyu Li [a,c], Yilan Liao [d], Xiaoying Zheng [d]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China
[d] Institute of Population Science, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 1 November 2013
Accepted 6 December 2013
Available online 12 December 2013

Keywords:
Rough fuzzy set
Rule extraction
Spatial analysis
Attribute reduct
Fuzzy decision rule

ABSTRACT

With the development of data mining and soft computing techniques, it becomes possible to automatically mine knowledge from spatial data. Spatial rule extraction from spatial data with uncertainty is an important issue in spatial data mining. Rough set theory is an effective tool for rule extraction from data with roughness. In our previous studies, Rough set method has been successfully used in the analysis of social and environmental causes of neural tube birth defects. However, both roughness and fuzziness may co-exist in spatial data because of the complexity of the object and the subjective limitation of human knowledge. The situation of fuzzy decisions, which is often encountered in spatial data, is beyond the capability of classical rough set theory. This paper presents a model based on rough fuzzy sets to extract spatial fuzzy decision rules from spatial data that simultaneously have two types of uncertainties, roughness and fuzziness. Fuzzy entropy and fuzzy cross entropy are used to measure accuracies of the fuzzy decisions on unseen objects using the rules extracted. An example of neural tube birth defects is given in this paper. The identification result from rough fuzzy sets based model was compared with those from two classical rule extraction methods and three commonly used fuzzy set based rule extraction models. The comparison results support that the rule extraction model established is effective in dealing with spatial data which have roughness and fuzziness simultaneously.

1. Introduction

Spatial data mining is the process of discovering the interesting and previously unknown, but potentially useful, patterns and rules (association and classification rules) from spatial datasets [39,50]. Over the last two decades, spatial data mining has been widely used in many applications, such as categorizing and localizing the human action(s) contained in a video [30], evaluating the structural and topological consistency among multiple representations of complex regions with broad boundaries [10], mining the frequent trajectory patterns in a spatial–temporal database [21], extracting the spatial association rules from a remotely sensed image database [22], generating the appropriate polygons from heterogeneous spatial information [38], and analyzing the change of land use [9].

Extraction of spatial decision rules is one of the main targets of spatial data mining [39,40,50] and has been used in many real applications. Ester et al. [13] used ID3 to extract decision rules from spatial databases via the so-called generalized attributes which take the neighborhood relation into consideration through a predefined position path of neighbors. Pontius et al. [35] proposed a model to select the locations of land-use change by the decision rules generated using the nearest neighbors. Daniels [8] introduced domain knowledge base which consists of decision rules to help classification of land cover types of remotely sensed imagery. Frank et al. [15] took the spatial and aggregation literals, such as perimeter, location and area, of spatial objects into account when mining rules from spatial data via a Voronoi-based approach to take non-spatial features into the rule extraction process. Zhu and Hu [54] extracted rules by using support vector machines, which is originally difficult to explain to users or to be understand by users as a black box, via analyzing the consistent regions formed by samples in terms of classification boundary.

In the studies mentioned above, both condition attributes and decision attributes, which are used to describe the objects in a data table, are crisp. The extracted rules are also crisp, i.e., the antecedent and consequent of a rule are expressed in some accurate way. However, because of the complexity of the object world, the subjective limitation of human knowledge and the uncertainty

* Corresponding author. Tel./fax: +86 10 64888053.
E-mail address: gey@lreis.ac.cn (Y. Ge).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.12.008

intrinsic of spatial data, there exists fuzziness in the representations of geographical phenomena [40] such as the classification of transition zones of land cover types, and the detection of the influence of environmental factors on the incidence rate of birth defects [4]. If the traditional models are used in extracting rules from spatial data with fuzzy descriptions by nature, the fuzzy concepts or fuzzy decisions should be degraded into crisp ones. For example, the transition zones between grassland and forest should be assigned to one specific category. The degradation may lead to information loss and decreases prediction accuracy. Meanwhile, the rules extracted using classical models will have crisp decision. These rules are less explainable than the fuzzy decision rules [12]. Some researchers then used fuzzy set theory to handle this issue. For example, Hu et al. [17] proposed the fuzzy-grid-based rule mining algorithm to generate fuzzy association rules. Niu et al. [31] used fuzzy concept lattice to mine the spatial association rules. In fact, there also exists another kind of uncertainty, i.e., roughness, in the geographical data tables.

In general, roughness is interpreted as the uncertainty of a concept while the concept cannot be precisely expressed with other concepts. In data analysis field, roughness is firstly studied by Pawlak in 1982 [34], where a concept is defined as an object subset of the universe with some property. In rough set theory, two operators, upper and lower approximations, are designed to roughly depict a target concept. A concept is called to be rough when its boundary region, i.e., the difference between the upper and lower approximations, is nonempty. The roughness of a concept derives from its boundary region. Especially, in geographical phenomena, roughness means that a target concept cannot be precisely described by the available information granules formed by spatial objects' features. For example, the spectral information of the remotely sensed imagery may not precisely describe a landcover type in the study area. As the roughness concept can be precisely defined when more additional features are collected, it is not a fuzzy concept by nature. Accordingly, it is inappropriate to use fuzzy sets to handling roughness. For example, rock and soil are hard to be discerned using existing band in remotely sensed imagery. But they may be clearly distinguished using other unavailable bands. Rock and soil do not have ill definition of boundary in such situation. Accordingly, fuzzy sets based model is not suitable for such situation. This means that the existing methods for rule extraction need to be extended for well working in spatial data analysis.

Rough set theory [34] is an effective tool in dealing with roughness and it can be used in extracting decision rules in spatial data with roughness [6,7]. However, the classical rough set theory is only suitable to the cases that objects are described by the nominal type of condition attributes and the crisp type of decision attributes in a data table. Nonetheless, in many real applications, the decision value of an object is fuzzy. Taking Heshun Neural Tube Birth Defects (NTD) data as an example, Bai et al. [3] used rough set theory to extract spatial decision rules from Heshun NTD data. However, the decision attribute in that work is "whether there are NTD instances in a village", i.e., it is a Yes–No type decision attribute, which cannot reflect the severity degree of NTD for each village. For example, a village within ten instances suffers more than that within only one instance. It is obvious that treating NTD birth defects as a fuzzy concept on the domain of discourse of all villages in Heshun is a better way than using a Yes–No type decision attribute. It can more intuitively describe the severity degree of NTD birth defects in villages. By the discussion above there are two kinds of uncertainties, i.e., roughness and fuzziness, needed to be handled in this case. Two extensions of rough set theory provide tools for handling data of this type.

Rough fuzzy sets and fuzzy rough sets [11] were proposed to extend the classical rough set theory [34] which allows the existence of fuzziness in decision attributes. They combine the advantages of rough sets and fuzzy sets [52]. The difference between fuzzy rough sets and rough fuzzy sets is that fuzzy rough sets is designed for the cases that both conditional attribute and decision attribute are fuzzy sets of the universe while rough fuzzy sets are specialized in dealing with data tables with crisp conditions and fuzzy decisions. Although rough fuzzy sets are special cases of fuzzy rough sets, the modeling process of fuzzy rough sets needs the fuzzification of conditional attributes. The fuzzification process involves the selection of the fuzzification methods, which will increase the modeling complexity and introduce new source of uncertainty. Therefore, the fuzzy rough sets based model cannot completely replace rough fuzzy sets based model.

Many researches on modeling spatial data by using rough set theory have been reported such as the classification of remotely sensed imagery, modeling spatial topology between spatial objects, uncertainty analysis and rule extraction [2,3,22,33,45,51], while rough fuzzy sets and fuzzy rough sets attract little attention in spatial data mining. Ahlqvist et al. [1] defined rough fuzzy classification and proposed various kinds of accuracy measures on rough fuzzy classification. This model assesses the classification accuracy using some goodness functions that reflect the features of the classification result from different perspectives. However, it did not give the accuracy assessment of prediction result in terms of reference. Furthermore, little research seems to address the detailed process of the reduct which is an essential procedure in spatial data mining using rough fuzzy sets.

In this paper, we focus on the problem of rule extraction from spatial data with crisp condition attributes and fuzzy decisions. A rough-fuzzy set based rule extraction model is used to deal with both fuzziness and roughness in spatial data tables. Unlike other commonly used spatial rule extraction methods, this model can simultaneously consider roughness and fuzziness in data. This model firstly converts the spatial data into a fuzzy decision information system. Rough fuzzy sets are then used to find a reduction of the fuzzy decision information system. Next, some fuzzy decision rules are extracted from the reduced fuzzy decision information system. Using the extracted rules, unseen objects can be classified, and the classification result is assessed by using fuzzy entropy and fuzzy cross entropy. This model is used in the analysis of Heshun NTD data, which is a very critical issue in China and has been studied for years by the authors. The rough fuzzy classification results of NTD data is compared with the results based on two kinds of classical rule extraction methods and three fuzzy decision rule extraction methods. The experimental results show that rough fuzzy set is an appropriate model for spatial analysis with both roughness and fuzziness in data.

The present paper has the following organization. Section 2 outlines the concepts of fuzzy decision information system and rough fuzzy sets. Section 3 proposes a model based on rough fuzzy sets to extract spatial rules from fuzzy decision information system constructed from spatial data. An example of NTD data from Heshun, Shanxi, China is given in Section 4. Section 5 makes detailed discussion on the effectiveness of the new method via entropy based accuracy assessment and performs a comparison of the proposed model with other five commonly used rule extraction methods. The last section concludes this paper.

2. Fuzzy decision information systems and rough fuzzy sets

A decision information system is defined as a pair $(U, AT \cup D)$, where the universe U is a non-empty finite set of objects, AT is a non-empty finite condition attribute set which contains m_c elements, D is a non-empty finite decision attribute set which contains m_d elements ($m = m_c + m_d$), and $AT \cap D$ is empty. Any $a \in AT$