International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Neighborhood based decision-theoretic rough set models

Weiwei Li [a,b], Zhiqiu Huang [a], Xiuyi Jia [c,*], Xinye Cai [a]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[b] College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[c] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 1 April 2015
Received in revised form 22 October 2015
Accepted 6 November 2015
Available online xxxx

Keywords:
Neighborhood relation
Decision-theoretic rough set model
Attribute reduction

ABSTRACT

As an extension of Pawlak rough set model, decision-theoretic rough set model (DTRS) adopts the Bayesian decision theory to compute the required thresholds in probabilistic rough set models. It gives a new semantic interpretation of the positive, boundary and negative regions by using three-way decisions. DTRS has been widely discussed and applied in data mining and decision making. However, one limitation of DTRS is its lack of ability to deal with numerical data directly. In order to overcome this disadvantage and extend the theory of DTRS, this paper proposes a neighborhood based decision-theoretic rough set model (NDTRS) under the framework of DTRS. Basic concepts of NDTRS are introduced. A positive region related attribute reduct and a minimum cost attribute reduct in the proposed model are defined and analyzed. Experimental results show that our methods can get a short reduct. Furthermore, a new neighborhood classifier based on three-way decisions is constructed and compared with other classifiers. Comparison experiments show that the proposed classifier can get a high accuracy and a low misclassification cost.

1. Introduction

Pawlak rough set [32,33] is a very useful mathematical tool for knowledge representation, especially in describing the uncertainty of the data. In Pawlak rough set model, an object x will be classified into the category X if the equivalence class of x: $[x]$ is a subset of X, which means $p(X|[x]) = \frac{|X \cap [x]|}{|[x]|} = 1$. Therefore, Pawlak rough set model can be seen as a qualitative model. However, two kinds of limitations exist in this model as follows. One is the probability $p(X|[x])$ must be equal to 1, which is sensitive to noisy data. The other one is that the equivalence class $[x]$ is defined based on the indiscernibility relation, which is not capable of dealing with numerical data directly.

For the first limitation, decision-theoretic rough set model (DTRS) [44,45] introduces a generalized framework to solve it. It is well known that, without considering the tolerance of classification error, in Pawlak rough set model it is difficult to obtain an effective result from the data with noise. To overcome this problem, many researchers studied different probabilistic rough set (PRS) models [34,37,45,60]. Compared to Pawlak rough set model, all PRS models take the tolerance of classification error into account. $p(X|[x]) \geq \alpha$ is used to classify the object x into the category X, while α is a threshold between 0 and 1.

* Corresponding author.
E-mail addresses: liweiwei@nuaa.edu.cn (W. Li), zqhuang@nuaa.edu.cn (Z. Huang), jiaxy@njust.edu.cn (X. Jia), xinye@nuaa.edu.cn (X. Cai).

http://dx.doi.org/10.1016/j.ijar.2015.11.005
0888-613X/

As a generalized rough set model, DTRS provides a unified and comprehensive framework for interpreting and determining the required thresholds. Based on minimum Bayesian decision cost procedure, DTRS can compute the required thresholds from given cost functions. Different thresholds for different PRS models can be deduced from appropriate cost functions.

Another main contribution of DTRS to rough set theory is the introducing of the notion of three-way decisions [48,49]. Compared to classical two-way decisions in classification or decision problems, three-way decisions bring in a deferment decision based on acceptance decision and rejection decision. In rough set theory, category X is described by three regions. An object x will be classified into one of these regions. Three different classification results can be interpreted by three-way decisions. Acceptance decision classifies x into the positive region of X, deferment decision classifies x into the boundary region of X, and rejection decision classifies x into the negative region of X.

The advantage of DTRS is that it can deal with noisy data by considering the tolerance of classification error. Three-way decisions framework based on DTRS can convert some potential being misclassified objects into the boundary region for further examination, which means DTRS usually has a higher classification accuracy. However, the disadvantage of DTRS is that it cannot deal with numerical data directly, which is also another limitation of Pawlak rough set model. To overcome this problem, many studies usually adopt two kinds of methods in real applications. One is discretization [7], the numerical data is discretized before applying rough set models. Each discretized interval is seen as a nominal value of the attribute. The other one is defining equivalence class based on other relations instead of indiscernibility relation, such as distance functions [24], fuzzy binary relations [42,56], dissimilarity and similarity measures [40], which can be used to measure and represent the continuous values. Generally, all these relations can be understood as special classes of neighborhood systems [26,46].

In general, there exist many data which contain numerical values and noisy values simultaneously. In this regard, we propose an extended rough set model, i.e. neighborhood based decision-theoretic rough set model (NDTRS) to deal with this type of numerical data which are accompanied by noisy values. Besides introducing some basic concepts which are usually defined in rough set models, two kinds of attribute reducts are also defined. The first attribute reduct keeps the positive region of the decision table unchanged or extended. The second attribute reduct can minimize its induced decision cost. Heuristic approaches to computing positive region preservation based attribute reduct and minimum cost attribute reduct are designed. Compared to supervised or unsupervised discretization methods, the proposed heuristic approaches can get the shorter reducts. A three-way decisions based neighborhood classifier is also proposed for classification problem. Compared to the classical two-way decisions based neighborhood classifier and several classical classifiers including C4.5, k-NN and SVM, our proposed classifier can obtain a higher accuracy and a less misclassification cost on several data sets.

We organize the paper as follows: Section 2 gives a briefly introduction about the related work. Section 3 introduces the framework of NDTRS. Section 4 defines two kinds of attribute reducts in NDTRS. Section 5 gives a three-way decisions based neighborhood classifier for classification problem. Section 6 presents the experimental results. Section 7 concludes this paper.

2. Related work

In this section, we will briefly introduce the related work on DTRS and neighborhood systems.

In recent years, DTRS has attracted much attention [31]. On the extension of the model, the relationships between DTRS and other PRS models have been investigated by Yao [47]. Many other extended models have been proposed based on DTRS, including multi-view DTRS [58], multiple-category DTRS [28,29,57], game-theoretic rough set model [8], Naive Bayesian DTRS [52], multi-granulation DTRS [35] and triangular fuzzy DTRS [25]. Attribute reduction in DTRS is also thoroughly discussed by many researchers [14,15,23,50]. On the application of the model, DTRS has been successfully applied in many different areas, including software defect prediction [20], text classification problem [21], clustering problem [53,22], spam filtering problem [16,59], and government decision analysis [30]. It is worth mentioning that a fuzzy DTRS approach was proposed to deal with real-valued data with fuzzy relation information [56]. We do not compare their method in this paper because we do not have any fuzzy relation information as the domain knowledge on the data sets compared in our experiments.

In the studies of neighborhood systems [9,10,18,39,43,54], using distance functions is a very common and useful method. Hu et al. [11] adopted three kinds of distance functions and proposed a neighborhood based rough set model, which is easy to understand and implement. On the extension of the model, Lin et al. [27] developed a neighborhood based multigranulation rough set in the framework of multigranulation rough sets. Wu and Zhang [43] investigated properties of neighborhood operator systems and rough set approximation operators. For attribute reduction and classification in neighborhood systems, Du et al. [6] discussed the attribute reduction and the rule learning based on a neighborhood covering space. Chen et al. [2] set up a connection between neighborhood-covering rough sets and evidence theory to establish a basic framework of numerical characterizations of attribute reduction. Hu et al. [10] defined a novel feature evaluation measure for feature selection in neighborhood rough set model. They also proposed large-margin nearest neighbor classifiers via sample weight learning [12]. (See Table 1.)

Please cite this article in press as: W. Li et al., Neighborhood based decision-theoretic rough set models, International Journal of Approximate Reasoning (2015), http://dx.doi.org/10.1016/j.ijar.2015.11.005