Applied Soft Computing 38 (2016) 942–954

Contents lists available at ScienceDirect

Applied Soft Computing

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

Three-way decision reduction in neighborhood systems

Yumin Chen [a,b,*], Zhiqiang Zeng [a], Qingxin Zhu [b], Chaohui Tang [a]

[a] Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China
[b] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

ARTICLE INFO

Article history:
Received 7 February 2015
Received in revised form 15 August 2015
Accepted 8 October 2015
Available online 2 November 2015

Keywords:
Rough set theory
Three-way decisions
Attribute reduction
Neighborhood systems

ABSTRACT

Rough set reduction has been used as an important preprocessing tool for pattern recognition, machine learning and data mining. As the classical Pawlak rough sets can just be used to evaluate categorical features, a neighborhood rough set model is introduced to deal with numerical data sets. Three-way decision theory proposed by Yao comes from Pawlak rough sets and probability rough sets for trading off different types of classification error in order to obtain a minimum cost ternary classifier. In this paper, we discuss reduction questions based on three-way decisions and neighborhood rough sets. First, the three-way decision reducts of positive region preservation, boundary region preservation and negative region preservation are introduced into the neighborhood rough set model. Second, three condition entropy measures are constructed based on three-way decision regions by considering variants of neighborhood classes. The monotonic principles of entropy measures are proved, from which we can obtain the heuristic reduction algorithms in neighborhood systems. Finally, the experimental results show that the three-way decision reduction approaches are effective feature selection techniques for addressing numerical data sets.

1. Introduction

Feature reduction is a quite useful data preprocessing technique, aiming to determine a minimal feature subset from a problem domain while retaining a suitably high classification accuracy of decision systems. It is also mentioned as a semantic-preserving dimension reduction [15,7], attribute reduction [47] and feature selection [8], applied in many areas including pattern recognition [36], machine learning [46], data mining [26] and big data [32], etc. As to the machine learning problems, feature reduction is an important preprocess to achieve the essence by deleting noisy, irrelevant or misleading features.

Rough set theory (RST) [29], proposed by Pawlak, has been used successfully as a reduction tool to discover data dependencies by reducing the redundant features that contained in a data set. In the last two decades, many methods for feature reduction have been developed in researches of rough set theory [30,33,28,34,18,38]. According to the different knowledge representations, methods of rough set reduction are mainly classified into two categories: algebra methods and information entropy methods [27]. The methods of algebra representation include positive region approaches and discernibility matrix approaches. In the algebra representations, a reduct is defined by a positive region preservation or a discernibility matrix function. Pawlak in [30] proposed a reduct of positive region preservation that does not vary the positive region or the quality of classification. Skowron and Rauszer in [33] introduced a new approach to knowledge reduction by providing a function of discernibility matrix that can obtain all reducts of a given data set. In the information entropy representations, a reduct is defined by Shannon's entropy and its extensions. Miao and Hu in [28] proposed an entropy measure for information systems and a mutual information measure for decision systems. Furthermore, they presented two heuristic reduction algorithms based on the information entropy and mutual information respectively. Slezak in [34] discussed the applications of feature reduction in data mining area by an entropy measure and extracted decision rules from big data sets. Liang et al. in [18] introduced a new information entropy to incomplete data reduction process for measuring the uncertainty of incomplete information systems. Wang et al. in [38] developed a novel reduction algorithm based on the condition entropy of a decision system.

Since the real world existing massive uncorrect, uncertain and noisy data, the Pawlak rough set theory is extended by introducing probabilistic theory. The three-way decision theory proposed by Yao in [44,45] comes from Pawlak rough set theory and probabilistic theory [24,27,40,41]. Its main purpose is to interpret the positive, negative and boundary regions of rough sets as three decisions outcomes: acceptance, rejection and uncertainty (or deferment) in a ternary classification respectively [9]. In recent years, many researchers mainly focus on model extensions and practical applications of three-way decisions. The one category is model researches on three-way decisions. It mainly contains the extension models of rough sets, such as decision-theoretic rough sets [2,42], variable precision rough sets [17,52], Bayesian rough sets [35], fuzzy rough sets and rough fuzzy sets [3], interval-valued fuzzy rough sets [5,10] and dominance-based fuzzy rough sets [4]. The another category is practical applications of three-way decisions, such as government decisions [25], text classification [19], information filtering [20], email filtering [16], investment decisions [23], cluster analysis [22] and attribute reduction of three-way decisions [27].

However, the Pawlak rough set reduction and three-way decision reduction are also established on the equivalence approximate space and only compatible for categorical data sets. They need to scatter the records when processing continuous numerical data, this will lead to losing of information (including the neighborhood structure information and order structure information in real spaces) [15,11], so the reducts of numerical data sets are strongly related with the methods of scattering. To overcome this drawback, many extensions of Pawlak rough set theory and

* Corresponding author at: Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China. Tel.: +86 13799762651.
E-mail address: cym0620@163.com (Y. Chen).

http://dx.doi.org/10.1016/j.asoc.2015.10.059
1568-4946/

their corresponding definitions on attribute reduction have been presented, such as fuzzy rough sets [7,43,12], tolerance approximate models [31], similarity rough approximate models [37], dominance approximation relation models [6], covering approximation models [49–51] and neighborhood granular models [21,39]. Among all the extensions, the neighborhood rough set model [13,14] can be regarded as a specified implementation of the neighborhood granular model. The neighborhood rough set model can process both numerical and categorical data sets via the δ-neighborhood set, which will not break the neighborhood structure and order structure of data sets in real spaces.

However, there are some premature theories and disadvantages about three-way decision reduction. The monotonicity of decision regions (positive region, boundary region or negative region) preservation reductions no longer holds in the three-way decision systems. The monotonicity is very important for constructing heuristic reduction algorithms, which can quickly converge to be an attribute reduct. Monotonously heuristic information that can guide searches for the attribute reduction in the three-way decision model is unavailable. Therefore, monotonous measure functions need to be developed to design heuristic reduction algorithms in the three-way decision model. In addition, few studies have been published on three-way decision reduction in neighborhood systems. In this paper, we propose a novel method for the attribute reduction of neighborhood decision systems based on three-way decision theory. We review some concepts related to attribute reductions with neighborhood rough sets and focus on the definitions of attribute reducts based on decision region preservation of three-way decisions in neighborhood decision systems. As monotonously heuristic information to design attribute reduction algorithms of the three-way decisions in neighborhood systems is lacking. Hence, we propose three monotonic measure functions by considering variants of the condition information entropy. Finally, heuristic computing methods of three types of decision region preservation reducts are also given.

The rest of this paper is structured as follows. Section 2 describes Pawlak rough set theory, three-way decision theory and the neighborhood rough set model. Section 3 introduces the positive region preservation reduct, the boundary region preservation reduct and the negative region preservation reduct based on a three-way decision model in neighborhood systems, and gives the computing methods for three types of decision region preservation reducts by introducing three novel monotonic measures. Furthermore, we develop heuristic algorithms to obtain three types of decision region preservation reducts. Section 4 provides an experimental analysis, including the theoretic analysis and the effectiveness of the proposed attribute reduction methods. Section 5 concludes the paper, and proposes further work in this area.

2. Preliminaries

In this section, we recall the basic notions related to Pawlak rough sets [30], three-way decisions [44,47] and neighborhood rough sets [14].

2.1. Pawlak rough sets and a Pawlak reduct

Definition 1. [29] A decision system is a four-tuple: $S=(U, C\cup D, V, f)$, where $U=\{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects called universe, C is a non-empty finite set of condition attributes, D is a finite set of decision attributes, $C\cap D=\emptyset$; V_a is a non-empty set of values of $a \in (C\cup D)$, and $f: U\times(C\cup D)\to V_a$ is an information function that maps an object in U to exactly one value in V_a.

For brevity, a decision system is denoted by $S=(U, C\cup D)$.

Definition 2. [29] Given a decision system $S=(U, C\cup D)$, for a subset $B\subseteq C$, an indiscernibility relation is defined by:

$$IND(B)=\{(x,y)\in U\times U|\forall b\in B, f(x,b)=f(y,b)\}. \tag{1}$$

Obviously, $IND(B)$ is an equivalence relation, which is reflexive, symmetric and transitive. The family of all equivalence classes of $IND(B)$ will be denoted by $U/IND(B)$, or simply U/B; an equivalence class of $IND(B)$ containing x will be denoted by $[x]_B$.

Definition 3. [29] Given a decision system $S=(U, C\cup D)$, for an attribute subset $B\subseteq C$ and an object subset $X\subseteq U$, the lower and upper approximations of X with respect to B are defined by:

$$B_*(X)=\{x\in U|[x]_B\subseteq X\}=\cup\{[x]_B|[x]_B\subseteq X\}; \tag{2}$$

$$B^*(X)=\{x\in U|[x]_B\cap X\neq\emptyset\}=\cup\{[x]_B|[x]_B\cap X\neq\emptyset\}. \tag{3}$$

The ordered pair $\langle B_*(X), B^*(X)\rangle$ is called a Pawlak rough set of X with respect to the equivalence relation $IND(B)$. According to the lower and upper approximations, one can obtain the positive, boundary and negative regions [29]:

$$POS_B(X)=B_*(X); \tag{4}$$

$$BND_B(X)=B^*(X)-B_*(X); \tag{5}$$

$$NEG_B(X)=U-POS_B(X)\cup BND_B(X)=U-B^*(X). \tag{6}$$

The positive region $POS_B(X)$ consists of all objects that are definitely contained in the set X. The negative region $NEG_B(X)$ consists of all objects that are definitely not contained in the set X. The boundary region $BND_B(X)$ consists of all objects that may be contained in X. Because approximations are from equivalence classes, inclusion into the boundary region reflects uncertainty about the classification of objects.

A classical attribute reduct in Pawlak rough set model is a relative reduct with respect to the decision attribute D, which is defined by requiring that the positive region of the decision attribute D is unchanged.

Definition 4. [29] Given a decision system $S=(U, C\cup D)$, and $U/D=\{D_1, D_2, \ldots, D_n\}$, an attribute set $R\subseteq C$ is a Pawlak reduct of C with respect to D if it satisfies the following two conditions:

(1) $POS_R(D)=POS_C(D)$;
(2) for any attribute $a\in R$, $POS_{R-\{a\}}(D)\neq POS_C(D)$,

where $POS_C(D)=\bigcup_{i=1}^{n}POS_C(D_i)$.

In this definition, condition (1) is also called a positive preservation condition and condition (2) is called a set independency condition.

2.2. Three-way decision rough set model

In this subsection, we introduce the basic concepts related to the three-way decision theory proposed by Yao in [44]. There are only two states and three actions (accept, defer and reject). The state set $\Omega=(X, \neg X)$ indicates that an element is in X and not in X, and the action set is $A=\{a_P, a_B, a_N\}$, where a_P, a_B and a_N represent the three actions of deciding that an object is in the sets $POS(X)$, $BND(X)$ and $NEG(X)$, respectively. Moreover, when an object belongs to X, let λ_{PP}, λ_{BP} and λ_{NP} denote the costs of taking the actions a_P, a_B and a_N, respectively; when an object does not belong to X, then let λ_{PN}, λ_{BN} and λ_{NN} denote the costs of taking the same three actions, respectively. The loss functions regarding the states X and $\neg X$ can be expressed by a 2×3 matrix, as follows:

	a_P	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
$\neg X$	λ_{PN}	λ_{BN}	λ_{NN}

Given a decision system $S=(U, C\cup D)$, for a subset $B\subseteq C$, $[x]_B$ denotes the equivalence class of x with respect to $IND(B)$, and the probabilities for the two complement states are denoted by $P(X|[x]_B)=\frac{|X\cap[x]_B|}{|[x]_B|}$ and $P(\neg X|[x]_B)=1-P(X|[x]_B)$.

From the matrix, the expected loss associated with taking different actions can be expressed by:

$$\Re(a_P|[x])=\lambda_{PP}P(X|[x]_B)+\lambda_{PN}P(\neg X|[x]_B);$$
$$\Re(a_B|[x])=\lambda_{BP}P(X|[x]_B)+\lambda_{BN}P(\neg X|[x]_B);$$
$$\Re(a_N|[x])=\lambda_{NP}P(X|[x]_B)+\lambda_{NN}P(\neg X|[x]_B).$$

The Bayesian decision procedure leads to the following minimum-risk decision rules: