Knowledge-Based Systems 71 (2014) 146–161

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Region-based quantitative and hierarchical attribute reduction in the two-category decision theoretic rough set model

Xianyong Zhang [a,b,c,*], Duoqian Miao [b,c]

[a] College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610068, PR China
[b] Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China
[c] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804, PR China

ARTICLE INFO

Article history:
Received 12 January 2014
Received in revised form 9 July 2014
Accepted 24 July 2014
Available online 7 August 2014

Keywords:
Rough set theory
Decision-theoretic rough set
Attribute reduction
Quantitative reduct
Hierarchical reduct

ABSTRACT

Quantitative attribute reduction exhibits applicability but complexity when compared to qualitative reduction. According to the two-category decision theoretic rough set model, this paper mainly investigates quantitative reducts and their hierarchies (with qualitative reducts) from a regional perspective. (1) An improved type of classification regions is proposed, and its preservation reduct (CRP-Reduct) is studied. (2) Reduction targets and preservation properties of set regions are analyzed, and the set-region preservation reduct (SRP-Reduct) is studied. (3) Separability of set regions and rule consistency is verified, and the quantitative and qualitative double-preservation reduct (DP-Reduct) is established. (4) Hierarchies of CRP-Reduct, SRP-Reduct, and DP-Reduct are explored with two qualitative reducts: the Pawlak-Reduct and knowledge-preservation reduct (KP-Reduct). (5) Finally, verification experiments are provided. CRP-Reduct, SRP-Reduct, and DP-Reduct expand layer by layer Pawlak-Reduct and exhibit quantitative applicability, and the experimental results indicate their effectiveness and hierarchies regarding Pawlak-Reduct and KP-Reduct.

1. Introduction

Rough set theory (RS-Theory) [1,2] is a novel mathematical theory for uncertainty descriptions and an important applicable methodology for knowledge discovery. In particular, it can effectively process uncertain, imprecise, and incomplete information. Thus, the model-based uncertainty description and reduction-based knowledge discovery become its two main issues, where the qualitative mechanism-based quantitative extension plays an increasingly important role.

The classical Pawlak-Model [1,2] is qualitative and thus has accuracy. However, the qualitative absoluteness can also cause some limitations and problems, such as over-fitting. In fact, Pawlak-Model cannot fully capture latent useful knowledge in the uncertainty boundary. In contrast, quantitative models resort to some measures and thresholds to express quantization approximation and fault tolerance, so they can tackle data sets with noises, thus holding important application significance; moreover, they usually conduct theoretical expansion for qualitative Pawlak-Model. Thus, the probabilistic rough set (PRS) [3–6] utilizes the probability uncertainty measure to exhibit application merits regarding measurability, generality, and robustness, and it includes several concrete models, such as the decision-theoretic rough set (DTRS) [7] and variable precision rough set (VPRS) [8].

Attribute reduction is a fundamental subject in RS-Theory due to its optimization and generalization for data mining. The classical reduction is related to Pawlak-Model and thus reflects a qualitative approach, and different reduction algorithms were extensively explored in [9–14]. In contrast, quantitative reduction mainly utilizes the quantitative mechanisms and advantages to achieve deep development and extensive applications; for example, Refs. [15–26] studied DTRS-Reduction and VPRS-Reduction, respectively. For the decision table, the classical reduction theory mainly depends on the classification-positive region (C-POS). Thus, Pawlak-Reduction directly preserves C-POS due to the change monotonicity of qualitative C-POS. However, quantitative region exhibits the change non-monotonicity, and quantitative reduction usually accompanies some anomalies [15,24,25]. In fact, Ref. [27] verified that quantitative regions have the essential change uncertainty, which determines the change non-monotonicity. Thus, quantitative reduction has already transcended qualitative Pawlak-Reduction and thus becomes a complex problem. For this

* Corresponding author at: College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610068, PR China.
E-mail addresses: xianyongzh@sina.com.cn (X. Zhang), miaoduoqian@163.com.cn (D. Miao).

http://dx.doi.org/10.1016/j.knosys.2014.07.022
0950-7051/

difficulty, we aim to conduct some systematical studies by virtue of a concrete quantitative model.

PRS usually needs thresholds for quantitative applications, so threshold determination becomes a critical task. In particular, DTRS achieves thresholds' semantics and calculation by using the Bayesian risk decision and three-way decision semantics [7]; moreover, DTRS also establishes a basic platform for quantitative explorations via its expansion and representativeness. For DTRS, Refs. [28,29] analyzed three-way decisions and their superiority, Refs. [30–33] discussed model development and threshold calculation, Refs. [34–38] researched model applications (regarding regression, clustering, and semi-supervised leaning), Refs. [38–41] exploited multi-category construction. For DTRS-Reduction, Ref. [15] proposed general reducts by mining transcendental measures for the dependency degree; moreover, Refs. [16–20] summarized the existing methods, including the positive-based reduct, nonnegative-based reduct, cost-based reduct, and distribution-based reduct.

Against the above backgrounds, quantitative reduction exhibits applicability but complexity, and DTRS is a fundamental PRS and its attribute reduction can reflect some essence of quantitative reduction. Thus, this paper concentrates on DTRS-Reduction in the decision table. Note that the two-category case corresponds to the fundamental issue for DTRS, and it is also linked to a usual classification task in the decision table. In fact, it causes relatively clear regional structure for RS-Theory by complementary simplification, thus underling multiple-category generalization; moreover, it can also provide some verification analyzes by degeneration. Therefore, our discussion is mainly within the two-category framework, and this restriction becomes a rational strategy in view of the complexity of quantitative reduction. In particular, granular computing (GrC) [42,43] emphasizes multiple levels and provides a structural approach for hierarchical information processing, and Refs. [44–48] conducted GrC studies for RS-Theory. Based on the GrC technology, we will construct hierarchical regional targets to systematically investigate hierarchical DTRS-Reduction on a basic premise of reduction expansion.

According to the two-category DTRS-Model, this paper mainly investigates quantitative reducts and their hierarchies (with qualitative reducts) from a regional perspective. It involves the following five parts. (1) An improved type of classification regions is proposed, and its preservation reduct (CRP-Reduct) is studied. (2) Reduction targets and preservation properties of set regions are analyzed, and the set-region preservation reduct (SRP-Reduct) is studied. (3) Separability of set regions and rule consistency is verified, and the quantitative and qualitative double-preservation reduct (DP-Reduct) is established. (4) Hierarchies of CRP-Reduct, SRP-Reduct, and DP-Reduct are explored with two qualitative reducts: the Pawlak-Reduct and knowledge-preservation reduct (KP-Reduct). (5) Finally, verification experiments are provided. In summary, the main contribution of our works is to construct three types of quantitative reducts and to further investigate their hierarchies with two types of qualitative reducts, and structural regions act as a main perspective in view of the two-category feature. As a result, CRP-Reduct, SRP-Reduct, and DP-Reduct expand layer by layer Pawlak-Reduct and exhibit quantitative applicability, and the experimental results indicate their effectiveness and hierarchies regarding Pawlak-Reduct and KP-Reduct.

The rest of this paper is organized as follows. Section 2 reviews basic models and reducts. Section 3 constructs an improved type of classification regions, and CRP-Reduct. Section 4 studies set-region preservation and SRP-Reduct. Section 5 discusses double-preservation and DP-Reduct. Section 6 investigates hierarchies of five reduction types. Section 7 conducts experimental analyzes. Finally, Section 8 concludes this paper.

2. Preliminaries

For simplification, abbreviations are first provided for several repeated terms. First alphabet-based replacement includes: Set → S, Classification → C, Region → R, Preservation → P, Double → D, and Knowledge → K.

(1) S-Region and C-Region denote the set region and classification region, respectively. Concretely, POS, BND, and NEG denote the set positive, boundary, negative regions, respectively, while C-POS, C-BND, and C-NEG denote the classification positive, boundary, negative regions, respectively.
(2) CR-Preservation, SR-Preservation, D-Preservation, and K-Preservation denote C-Region preservation, S-Region preservation, double preservation (of set regions and rule consistency), and knowledge preservation, respectively. Furthermore, CRP-Reduct, SRP-Reduct, DP-Reduct, and KP-Reduct denote corresponding preservation reducts.

Next, this section reviews Pawlak-Model, DTRS-Model, and their reducts.

2.1. Pawlak-Model and Pawlak-Reduct

Pawlak-Model and Pawlak-Reduct [1,2] are first reviewed.

U is a finite universe, $\mathscr{R}$ is a family of equivalence relations, and $(U,\mathscr{R})$ constitutes a knowledge base. Let $\emptyset \neq R \subseteq \mathscr{R}, \cap R$ determines an equivalence relation – $IND(R)$. Knowledge R refers to classified structure $U/IND(R)$ with granule $[x]_R$. Thus, (U,R) constitutes an approximate space, where set $X \subseteq U$ is also called a concept. In Pawlak-Model, the lower and upper approximations of X are defined by

$$\underline{apr}_R X = \{x|[x]_R \subseteq X\}, \overline{apr}_R X = \{x|[x]_R \cap X \neq \emptyset\}.$$

$$\begin{cases} \mathrm{POS}_R(X) = \underline{apr}_R X, \\ \mathrm{NEG}_R(X) = U - \overline{apr}_R X, \\ \mathrm{BND}_R(X) = \overline{apr}_R X - \underline{apr}_R X \end{cases} \tag{1}$$

further denotes POS, NEG, and BND.

The decision table (D-Table) is an important information table with classification tasks. In D-Table $(U, C \cup D), C$ and D include condition and decision attributes, respectively, and the decision rule is related to the function $d_x(a) = a(x)$, where $x \in U, a \in C \cup D$. D-Table is consistent, if all its decision rules are consistent, i.e., arbitrary decision rule d_x satisfies $d_x|C = d_y|C \Rightarrow d_x|D = d_y|D, \forall x \neq y$; otherwise, it is inconsistent. Moreover, condition attribute subset A determines an equivalence relation and knowledge, where $\emptyset \neq A \subseteq C$.

Definition 2.1 (*Qualitative Type*). In Pawlak-Model, C-Regions are qualitative and are composed of following C-POS and C-BND:

$$\begin{cases} \mathrm{POS}_A(D) = \bigcup\limits_{X \in U/IND(D)} \underline{apr}_{IND(A)} X, \\ \mathrm{BND}_A(D) = U - \mathrm{POS}_A(D), \end{cases} \tag{2}$$

$\mathrm{POS}_A(D)$ describes certain granules for classification. $\mathrm{POS}_{B'}(D) \subseteq \mathrm{POS}_B(D)$ if $B' \subseteq B \subseteq C$, so C-POS change has monotonicity. Thus, Pawlak-Reduct is naturally established by preserving C-POS; moreover, dependency degree $\gamma_A(D) = \frac{|\mathrm{POS}_A(D)|}{|U|}$ is important for evaluating classification quality.

Definition 2.2 (*Pawlak-Reduct*). B is Pawlak-Reduct of C, if it satisfies C-POS preservation and set independence, i.e.,