Knowledge-Based Systems 109 (2016) 84–89

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A unified reduction algorithm based on invariant matrices for decision tables

Guilong Liu*, Zheng Hua, Jiyang Zou

School of Information Science, Beijing Language and Culture University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 28 January 2016
Revised 20 June 2016
Accepted 21 June 2016
Available online 21 June 2016

Keywords:
Attribute reduction
Discernibility matrix
Decision table
Equivalence relation
Invariant matrix

ABSTRACT

Attribute reduction is an important issue for decision analysis in databases. Absolute reduction, distributive reduction and positive region reduction are the most common types of attribute reduction discussed in the existing literature. This paper considers these three reduction types from the viewpoint of matrices and proposes the concept of reduction invariant matrices for each type in decision tables. Based on invariant matrices, we establish a unified algorithm for all three reduction types in decision tables. We also study the relationships among the three reduction types. Finally, experiments with UCI data sets are presented to verify the effectiveness of the proposed algorithm.

1. Introduction

Attribute reduction is an important aspect of pattern recognition, decision analysis, and knowledge discovery in databases. It is an area of interest in databases and information systems research. Recall that an attribute reduction for a decision table reveals subsets of attributes that are jointly sufficient and individually necessary for preserving some particular properties. For example, attribute reduction is often needed to acquire brief decision rules from decision tables. In practical situations we need different types of attribute reduction to address different problems. Usually, the problems are proposed by users or domain experts. There are many types of attribute reduction [1,5–7,9,15,17,20,33,39] based on different criteria in the area of rough sets [18,19]. Jia, Shang, Zhou and Yao [8] gave a brief description of 22 existing reductions. They also answered the following three interesting questions: [8] (1) Why do we have so many different types of attribute reduction? (2) What are the differences among these reductions? (3) How to choose appropriate reductions for different users in different applications? Based on these three questions, they proposed a generalized attribute reduction approach to satisfy the needs of the users.

Pawlak [19], Skowron and Rauszer [24] were the first to propose the concept of positive region reduction for decision tables, where the reduction keeps the positive region unchanged. Their proposal also preserves deterministic rules with respect to decision attributes and is therefore applied in the extraction of deterministic decision rules from decision tables. This reduction type continues to be the most studied reduction in the rough set literature. Skowron and Rauszer [24] used a discernibility matrix to compute reduction sets. They also introduce absolute reduction, which keeps partitions unchanged, for information systems. Zhang, Mi and Wu [38] first studied distributive reduction, which keeps probability distribution unchanged, and established an efficient reduction algorithm for decision tables. The earliest studied attribute reductions are absolute reduction and positive region reduction.

Much work has been done on attribute reduction [25–29,34–36] for decision tables. Kryszkiewicz [11] extended the work of Skowron and Rauszer and defined the notion of the modified discernibility matrix and discernibility function for calculating reductions. Yao and Zhao [37] considered attribute reduction in decision-theoretic rough set (DTRS) models. Using representation of object subsets and indiscernibility relations in matrix forms, Luo, Li, Yi and Fujita [14] exploited matrix approaches to study an incremental DTRS approach for evolving data. Liu et al. [13] studied a quick attribute reduction algorithm for neighborhood rough set models. Chen et al. [4,30–32] discussed the reduction for relation decision systems and covering decision systems, and Liu et al. [12] extended their work to a more general relation decision system. Many authors [10,21,22] have also considered reduction problems for incomplete information systems. Recently, Miao and Lang [16] reviewed works on relative reductions in decision tables.

Absolute reduction, distributive reduction and positive region reduction are the most commonly studied reductions in rough set literature. This paper builds upon prior studies by considering

* Corresponding author.
E-mail address: liuguilong@blcu.edu.cn (G. Liu).

http://dx.doi.org/10.1016/j.knosys.2016.06.027
0950-7051/

the above three types of reduction for decision tables from the viewpoint of matrices. It is well known that different algorithms correspond to different reductions. Consequently, we wonder whether there exists a unified algorithm for different reductions. For each reduction, we define its invariant matrix, set up a unified reduction algorithm based on the invariant matrix, and partly answer the question affirmatively in this paper. We also study the relationships among the three reduction types.

The remainder of the paper is organized as follows. In Section 2, we review some basic concepts and properties of binary relations, decision tables and attribute reduction algorithms. In Section 3, we propose the concept of the reduction invariant matrices for the three reduction types. Based on invariant matrices, a unified discernibility matrix is established for a decision table. Section 4 studies the relationships among the three types of reduction. Section 5 designs an experimental process to verify our theoretical results. Finally, Section 6 concludes the paper.

2. Preliminaries

In this section, we recall some basic definitions and properties of binary relations and decision tables. Let U be a finite set of objects called the universal set. A binary relation R on U is a subset of $U \times U$. We use the notation xRy to denote that $(x, y) \in R$. Let R be a binary relation on U, recall that the left and right R-relative sets of an element, x, in U are defined as

$$l_R(x) = \{y | y \in U, yRx\} \text{ and } r_R(x) = \{y | y \in U, xRy\},$$

respectively. Recall the following terminology: R is called reflexive if $x \in r_R(x)$ for each $x \in U$; R is called symmetric if yRx whenever xRy for all $x, y \in U$; R is called transitive if whenever xRy and yRz, then xRz for all $x, y, z \in U$; and R is an equivalence relation if R is reflexive, symmetric and transitive. If R is an equivalence relation on U, then $[x]_R = r_R(x) = l_R(x)$ is the equivalent class of containing element $x \in U$. Suppose that $U = \{x_1, x_2, \ldots, x_n\}$ is a universal set with n elements and X is a subset of U, the characteristic function λ_X of X is the function from U to $\{0, 1\}$ such that $\lambda_X(x) = 1$ if $x \in X$, and $\lambda_X(x) = 0$ if $x \notin X$. Let R be a binary relation on U, relational matrix $M_R = (a_{ij})_{n\times n}$ of R is defined via $a_{ij} = \lambda_R(x_i, x_j)$ for $x_i, x_j \in U$.

An information system [19] is a pair (U, A), where U is a non-empty finite universal set and A is a non-empty finite set of attributes. Each $a \in A$ is a function a: $U \to V_a$, where V_a is the set of values of a, called the domain of a. If $A = C \cup D$ and $C \cap D = \emptyset$, then $(U, A) = (U, C \cup D)$ is called a decision table (or decision information system), C is condition attribute set and D is decision attribute set. For each subset $B \subseteq C$ and $B \neq \emptyset$, B induces an equivalence (or indiscernibility) relation R_B on U via $x_i R_B x_j$ if and only if $b(x_i) = b(x_j)$ for all $b \in B$ and $x_i, x_j \in U$. Thus each $a \in C \cup D$ can be viewed as an equivalence relation on U. If $R_C \subseteq R_D$, then $(U, C \cup D)$ is described as consistent; otherwise, $(U, C \cup D)$ is described as inconsistent. Additionally, every row of a decision table $(U, C \cup D)$ can be viewed as a decision rule. Clearly, the decision rule of x_i is consistent if and only if $[x_i]_{R_C} \subseteq [x_i]_{R_D}$. Moreover, if $[x_i]_{R_C} \subseteq [x_i]_{R_D}$, then the decision rule of x_i is deterministic; otherwise, the decision rule of x_i is uncertain. For the sake of simplicity, we always assume $D = \{d\}$ and abbreviate B for R_B in the sequel.

Many authors [1,5–7,9,15,17,20,33] have studied various types of attribute reduction with respect to different criteria in rough set literature. However, in this paper, we focus mainly on the three most common types of reduction.

Definition 2.1. Let $(U, C \cup D)$ be a relation decision table, $U/d = \{D_1, D_2, \ldots, D_s\}$ be the quotient set of D and $\emptyset \neq B \subseteq C$.

(1) If $\cap_{a\in C} a = \cap_{a\in B} a$, and for any subset $B' \subset B$, $\cap_{a\in C} a \neq \cap_{a\in B'} a$, then B is called a type-1(or absolute) reduction of C [23,24].

(2) Suppose that $\mu_{CD}(x_i) = (\frac{|[x_i]_C \cap D_1|}{|[x_i]_C|}, \ldots, \frac{|[x_i]_C \cap D_s|}{|[x_i]_C|})$ for $i = 1, 2, \ldots, n$ if $\mu_{CD}(u_i) = \mu_{BD}(u_i), i = 1, 2, \ldots, n$, and for any subset $B' \subset B$, $\mu_{CD}(x_i) \neq \mu_{B'D}(x_i)$ for some i, then B is called a type-2(or distributive) reduction of C [38].

(3) Suppose that $U_{CD} = \{x_i | [x_i]_C \subseteq [x_i]_D\}$, if $U_{CD} = U_{BD}$, and for any subset $B' \subset B$, $U_{CD} \neq U_{B'D}$, then B is called a type-3(or positive region) reduction of C [19].

Table 1
A decision table.

U	a_1	a_2	a_3	a_4	d
x_1	0	0	0	0	0
x_2	1	1	0	0	1
x_3	1	1	0	0	0
x_4	1	0	0	1	1
x_5	1	0	0	1	0
x_6	1	0	1	0	1
x_7	1	0	1	0	0
x_8	1	0	1	0	0

Generally, a reduction algorithm consists of the following three steps:

(1) Calculate the indiscernibility matrix.
(2) Construct the indiscernibility function f and transform f from its conjunctive normal form (CNF) into the disjunctive normal form (DNF).
(3) Obtain all minimal subset of attributes. That is, obtain a reduction of C.

Note that step (2) is usually time-consuming. Fortunately, Borowik and Luba [3] proposed a fast algorithm to solve this.

The three types of reduction correspond to three different discernibility matrices. The type-1 discernibility matrix is $F = (f_{ij})_{n\times n}$, where $f_{ij} = \{a | a \in C, a(x_i) \neq a(x_j)\}$ [24].

The type-2 discernibility matrix is $S = (s_{ij})_{n\times n}$ [34], where

$$s_{ij} = \begin{cases} \{a | a \in C, a(x_i) \neq a(x_j)\}, & \mu_{CD}(x_i) \neq \mu_{CD}(x_j) \\ \emptyset, & Otherwise \end{cases}.$$

The type-3 discernibility matrix is $T = (t_{ij})_{n\times n}$, [12] where

$$t_{ij} = \begin{cases} \{a | a \in C, a(x_i) \neq a(x_j)\}, & x_i \in U_{CD}, d(x_i) \neq d(x_j) \\ \emptyset, & Otherwise \end{cases}.$$

The following example illustrates that each type of reduction produces a different reduction result.

Example 2.1. Let $(U, C \cup D)$ be a decision table as shown in Table 1. $U = \{x_1, x_2, \ldots, x_8\}$, $C = \{a_1, a_2, a_3, a_4\}$ and $D = \{d\}$.

By direct computation, we obtain the following reduction results:

(1) All type-1 reductions are $\{a_1, a_2, a_3\}$, $\{a_1, a_2, a_4\}$, $\{a_1, a_3, a_4\}$, and $\{a_2, a_3, a_4\}$.
(2) All type-2 reductions are $\{a_1, a_3\}$, $\{a_1, a_2, a_4\}$, and $\{a_2, a_3, a_4\}$.
(3) All type-3 reductions are $\{a_1\}$, and $\{a_2, a_3, a_4\}$. Note that $\{a_2, a_3, a_4\}$ is a common reduction of the three types.

3. A unified reduction algorithm based on invariant matrices

In this section, we propose the concept of reduction invariant matrices for three types of reduction. For an equivalence relation R on U, suppose that $M_R = (a_{ij})_{n\times n}$ is the relational matrix of R. We