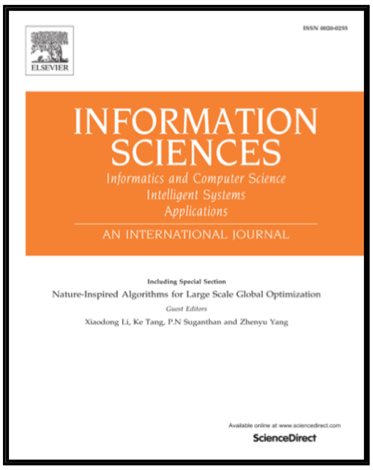

ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES

Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Including Special Section
Nature-Inspired Algorithms for Large Scale Global Optimization
Guest Editors
Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com
ScienceDirect

A novel attribute reduction approach for multi-label data based on rough set theory

Hua Li, Deyu Li, Yanhui Zhai, Suge Wang, Jing Zhang

PII: S0020-0255(16)30489-3
DOI: 10.1016/j.ins.2016.07.008
Reference: INS 12334

To appear in: *Information Sciences*

Received date: 2 November 2013
Revised date: 30 June 2016
Accepted date: 2 July 2016

Please cite this article as: Hua Li, Deyu Li, Yanhui Zhai, Suge Wang, Jing Zhang, A novel attribute reduction approach for multi-label data based on rough set theory, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.07.008

A novel attribute reduction approach for multi-label data based on rough set theory

Hua Li[a,b], Deyu Li[a,*], Yanhui Zhai[c], Suge Wang[c], Jing Zhang[c]

[a]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, Shanxi, China*
[b]*Department of Mathematics and Physics, Shijiazhuang Tiedao University, Shijiazhuang 050043, Hebei, China*
[c]*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China*

Abstract

Multi-label classification is an active research field in machine learning. Because of the high dimensionality of multi-label data, attribute reduction (also known as feature selection) is often necessary to improve multi-label classification performance. Rough set theory has been widely used for attribute reduction with much success. However, little work has been done on applying rough set theory to attribute reduction in multi-label classification. In this paper, a novel attribute reduction method based on rough set theory is proposed for multi-label data. First, the uncertainties conveyed by labels are analyzed, and a new type of attribute reduct is introduced, called complementary decision reduct. The relationships between complementary decision reduct and two representative types of attribute reducts are also investigated, showing significant advantages of complementary decision reduct in revealing the uncertainties implied in multi-label data. Second, a discernibility matrix-based approach is introduced for computing all complementary decision reducts, and a heuristic algorithm is proposed for effectively computing a single complementary decision reduct. Experiments on real-life data demonstrate that the proposed approach can effectively reduce unnecessary attributes and improve multi-label classification accuracy.

Keywords: Multi-label classification, Rough set theory, Attribute reduct,

[*] Corresponding author.
Email address: lidysxu@163.com (Deyu Li)

Complementary decision reduct, Discernibility matrix

1. Introduction

Multi-label data are omnipresent in real-world problems. In such data, one instance may be simultaneously associated with multiple labels. For example, an image may belong to multiple semantic classes, such as beach and mountain [1]; a piece of music may belong to more than one emotion class, such as happy-pleasant and relaxing-calm [41]; and a gene may be related to a set of functional classes, such as metabolism, transcription, and protein synthesis [5]. Within a multi-label classification framework, each instance is associated with a set of labels, and the task is to predict the unknown label sets of test instances by analyzing the known label sets of training instances.

As in traditional single-label classification problems, multi-label classification performance is strongly influenced by the quality of the input features (or attributes). Irrelevant or unnecessary features may lead to poor classification performance because the similarity between patterns from the same class may be reduced [45]. It is therefore desirable to reduce unnecessary features and select informative features to obtain more compact classification models and better generalization. Among various feature selection approaches, rough set theory, as a concrete granular computing model, has attracted much attention owing to the following advantages: its ability to discover data dependencies under the constraint of a limited collection of information granules, and its ability to reduce the number of attributes contained in a dataset using the data alone, without any additional information [27, 28, 29].

Feature selection in rough set theory is also called attribute reduction; it aims to remove unnecessary attributes while retaining the discernibility of objects under the original attributes. In the past few years, many types of attribute reduction approaches have been proposed according to various criteria [3, 4, 8, 11, 17, 22, 33, 42, 44, 47, 49, 52, 53]. For convenience, some of these techniques are briefly reviewed here. The positive region reduct, discussed by Grzymała-Busse in [9, 10], is a representative attribute reduction approach that aims to remove as many unnecessary attributes as possible while retaining the so-called positive regions, i.e., the consistency information. However, this reduct turns out to be too strict with respect to possible noise and fluctuations in data. A viable alternative that can overcome this restriction is the β-reduct, introduced by Ziarko in [53], which removes unnecessary attributes by allowing a controlled degree of inconsistency. Subsequently, Skowron used probabilistic tools to extend the rough set reduction laws and introduced the entropy reduct in his lectures at Warsaw University in 1993/1994. Following this line of thought, Ślęzak [39] proposed the approximate entropy reduct, which removes unnecessary attributes while approximately preserving decision information encoded in terms of information entropy. More recently, some attribute reduction methods in inconsistent systems and their relationships have been investigated by Kryszkiewicz [22], Li et al. [23], and Mi et al. [51]. In addition, because more than one reduct usually exists for a given dataset, calculation techniques for reducts have been widely discussed. For example, Skowron and Rauszer [38] developed a discernibility matrix-based approach to obtain all reducts. Unfortunately, it has been proved that finding all reducts, or finding an optimal reduct (i.e., a reduct with the minimum number of attributes), is an NP-hard problem [46]. Therefore, many algorithms have also been proposed to find a heuristic "optimal" reduct [12, 18, 43, 48]. For example, Hu and Cercone [13] proposed a heuristic attribute reduction algorithm for computing the positive region reduct; Wang et al. [43] used Shannon's conditional information entropy to construct a heuristic attribute reduction algorithm. However, because of their inefficient data structure and high computational cost, most existing attribute reduction algorithms cannot handle massive data well. Hence, several efficient strategies have been designed to improve the efficiency of a heuristic attribute reduction algorithm [16, 24, 31, 32, 36].

To the best of the authors' knowledge, however, little work has been done on applying rough set theory to attribute reduction in multi-label classification. Although directly applying existing attribute reduction methods to multi-label data is possible, it does not sufficiently take into account the uncertainties conveyed by labels and, therefore, could be enhanced further. In this paper, the uncertainties conveyed by labels are analyzed and a new type of attribute reduct is proposed, called complementary decision reduct. The relationships between complementary decision reduct and two representative types of attribute reducts are also investigated, showing significant advantages of complementary decision reduct in revealing the uncertainties of multi-label data. Furthermore, a discernibility matrix-based method is introduced for computing all possible complementary decision reducts, and a heuristic algorithm is proposed for effectively computing a single complementary decision reduct.

The rest of this paper is organized as follows. In Section 2, some basic notions of rough set theory are reviewed. Multi-label decision table is then introduced, and the limitations of directly applying existing attribute reduction methods to multi-label data are analyzed in Section 3. Section 4 introduces a new attribute reduct, referred to as complementary decision reduct; a discernibility matrix-based approach and a heuristic algorithm are also considered in this section. Section 5 reports a number of experimental results on several real-world multi-label datasets, and Section 6 concludes the paper.

2. Preliminaries

As a basis for further discussion, this section briefly reviews several basic concepts in rough set theory such as decision table, lower approximation, upper approximation, and positive region reduct.

In rough set theory, an information system with decision attributes is called a decision table, denoted by $S = (U, A \cup D)$, where $U = \{x_1, x_2, \cdots, x_n\}$ is a nonempty, finite set of objects; $A = \{a_1, a_2, \cdots, a_p\}$ is a nonempty, finite set of condition attributes; and $D = \{d_1, d_2, \cdots, d_s\}$ is a nonempty, finite set of decision attributes. In general, it is assumed that $A \cap D = \emptyset$ and that each attribute $a \in A \cup D$ forms a mapping $a : U \to V_a$, where V_a is the value domain of a.

Each nonempty subset $B \subseteq A$ determines an indiscernibility relation as follows:

$$R_B = \{(x, y) \in U \times U : \ a(x) = a(y), \ for\ all\ a \in B\}.$$

The indiscernibility relation R_B partitions U into a family of disjoint subsets given by $U/R_B = \{[x]_B : \ x \in U\}$, where $[x]_B$ denotes the equivalence class determined by x with respect to B, i.e.,

$$[x]_B = \{y \in U : \ (x, y) \in R_B\}.$$

Let $X \subseteq U$ and $B \subseteq A$. X can be characterized by a pair of lower and upper approximations:

$$\underline{R_B}(X) = \{x \in U : [x]_B \subseteq X\} = \bigcup\{[x]_B : [x]_B \subseteq X\},$$

$$\overline{R_B}(X) = \{x \in U : [x]_B \cap X \neq \emptyset\} = \bigcup\{[x]_B : [x]_B \cap X \neq \emptyset\}.$$

The lower approximation is called the positive region of X and can be denoted alternatively as $POS_B(X)$; X is called a rough set with respect to B if and only if $\underline{R_B}(X) \neq \overline{R_B}(X)$.

Attribute reduction is one of the most important topics in rough set theory, and numerous methods have been proposed according to various criteria. Among existing methods, positive region reduct, first discussed by Grzymała-Busse in [9] and [10], is a representative method.

Definition 1. *Let $S = (U, A \cup D)$ be a decision table, and let $B \subseteq A$. The subset B is a positive region reduct of S if and only if B satisfies the following conditions:*

(1) $POS_B(D) = POS_A(D)$,

(2) $POS_{B'}(D) \neq POS_A(D)$ for any $B' \subset B$,

where $POS_B(D) = \bigcup\limits_{i=1}^{r} POS_B(D_i)$ and $D_1, D_2, \cdots, D_r$ are decision classes, generated by the indiscernibility relation

$$R_D = \{(x, y) \in U \times U : d(x) = d(y),\ for\ all\ d \in D\}.$$

If B only satisfies condition (1), it is said that B is a positive region consistent set.

3. Multi-label data

This section first presents the definition of a multi-label decision table and then analyzes the limitations of applying existing attribute reduction approaches to multi-label data.

3.1. Multi-label decision table

Multi-label data can be represented as a multi-label decision table understood as a tuple $S = (U, A, L)$, where

- $U = \{x_1, x_2, \cdots, x_n\}$ is a nonempty finite set of objects;
- $A = \{a_1, a_2, \cdots, a_p\}$ is a nonempty finite set of condition attributes, called the condition attribute set; and
- $L = \{l_1, l_2, \cdots, l_q\}$ is a nonempty finite set of labels, called the label set.

Each condition attribute $a \in A$ forms a surjective function $a : U \to V_a$, where V_a is the value domain of a; each label $l \in L$ forms a surjective function $l : U \to V_l$, where $V_l = \{0, 1\}$ is the value domain of l. If the object x is associated with label l, then $l(x) = 1$; otherwise, $l(x) = 0$.

Some conventions in multi-label classification are as follows:

(1) The condition attribute set A and the label set L are disjoint, i.e., $A \cap L = \emptyset$.

(2) In multi-label classification, it is usually assumed that each object in U is associated with at least one label from the label set L [7, 20]. This means that unlabeled objects are irrelevant to multi-label classification and are not taken into account in this setting. Note that this convention is a prerequisite for the proposed approach, as discussed in Section 4.

(3) Each label from L is associated with at least one object in U [35].

The following example depicts a multi-label decision table in more detail:

Example 1. *A multi-label decision table $S = (U, A, L)$ is presented in Table 1, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$, $A = \{a, b, c\}$, and $L = \{l_1, l_2, l_3\}$. It can be seen that each object in U is associated with at least one label from L and that each label from L is associated with at least one object in U.*

Table 1: A multi-label decision table $S = (U, A, L)$

U	a	b	c	l_1	l_2	l_3
x_1	1	2	1	1	0	0
x_2	3	2	2	0	1	0
x_3	1	2	1	1	0	1
x_4	2	3	1	1	0	1
x_5	2	3	1	0	0	1
x_6	1	2	2	0	1	0
x_7	2	3	1	1	1	1
x_8	1	2	2	1	1	1
x_9	1	1	2	0	1	1
x_{10}	3	1	1	1	1	1
x_{11}	1	1	2	1	1	0

3.2. Limitations of existing attribute reduction approaches for multi-label data

This section is mainly devoted to analyzing the limitations of directly applying existing attribute reduction approaches to multi-label data.

For a multi-label decision table $S = (U, A, L)$, each label attribute can be viewed as a binary decision attribute, and then an indiscernibility relation R_L can be formed as follows:

$$R_L = \{(x, y) \in U \times U : \ l(x) = l(y), \ for\ all\ l \in L\}.$$

R_L partitions U into a family of mutually exclusive subsets given by $U/R_L = \{D_1, D_2, \cdots, D_r\}$, where $D_1, D_2, \cdots, D_r$ are decision classes. By taking into account the indiscernibility relation R_L and the corresponding decision classes, most existing attribute reduction approaches can be directly applied to multi-label data. As an example, the positive region reduct will be considered here, and the problem of deleting irrelevant or unnecessary condition attributes in a multi-label decision table will be addressed. The following example illustrates this process:

Example 2. *For the multi-label decision table $S = (U, A, L)$ given in Table 1, it can be calculated that*

$$\begin{aligned} U/R_A &= \{X_1, X_2, X_3, X_4, X_5, X_6\} \\ &= \{\{x_1, x_3\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_6, x_8\}, \{x_9, x_{11}\}, \{x_{10}\}\}, \\ U/R_L &= \{D_1, D_2, D_3, D_4, D_5, D_6, D_7\} \\ &= \{\{x_1\}, \{x_2, x_6\}, \{x_3, x_4\}, \{x_5\}, \{x_7, x_8, x_{10}\}, \{x_9\}, \{x_{11}\}\}. \end{aligned}$$

Hence, $POS_A(D) = \{x_2, x_{10}\} = X_2 \cup X_6$. This implies that the other equivalence classes X_1, X_3, X_4, and X_5 in U/R_A are all uncertain with respect to the label set L. For example, consider the equivalence class $X_1 = \{x_1, x_3\}$. Note that x_1 and x_3 are indiscernible with respect to A, whereas their respective label sets $\{l_1\}$ and $\{l_1, l_3\}$ are discernible with respect to L. This means that X_1 is uncertain with respect to the label set L. Furthermore, it can be determined that

$$\begin{aligned} U/R_{\{a,b\}} &= \{Y_1, Y_2, Y_3, Y_4, Y_5\} \\ &= \{X_1 \cup X_4, X_2, X_3, X_5, X_6\} \\ &= \{\{x_1, x_3, x_6, x_8\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_9, x_{11}\}, \{x_{10}\}\}, \\ U/R_{\{a,c\}} &= \{Z_1, Z_2, Z_3, Z_4, Z_5\} \\ &= \{X_1, X_2, X_3, X_4 \cup X_5, X_6\} \\ &= \{\{x_1, x_3\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_6, x_8, x_9, x_{11}\}, \{x_{10}\}\}. \end{aligned}$$

Because X_1, X_4 and X_5 are all uncertain with respect to L, they can be safely merged without any information loss. In other words, removing attribute c or b is valid from the perspective of rough set theory. Moreover, it is possible to determine that no more attributes can be removed from $\{a, b\}$ or $\{a, c\}$. Hence, both $\{a, b\}$ and $\{a, c\}$ are positive region reducts.

However, note that all objects in X_1 must be associated with label l_1 and may be associated with label l_3, and that all objects in X_4 must be associated with label l_2 and may be associated with labels l_1, l_3. Hence, the uncertainties of X_1 and X_4 are different, and the equivalence class Y_1, which is the union of X_1 and X_4, cannot preserve the uncertainties conveyed by labels. This implies that $\{a, b\}$ is not an appropriate attribute reduct.

By contrast, X_4 and X_5 share the certain label l_2 and the uncertain labels l_1 and l_3, and, hence, $Z_4 = X_4 \cup X_5$ preserves the uncertainties conveyed by labels. Therefore, the reduct $\{a, c\}$ is more valuable in reducing redundant attributes for multi-label data than $\{a, b\}$.

Through the above analysis, it is clear that some positive region reducts are not appropriate for multi-label data because they cannot preserve the uncertainties conveyed by labels. In fact, because the computation of positive region reduct must refer to the indiscernibility relation R_L, the uncertainties conveyed by labels are not thoroughly analyzed. In addition, note that most existing attribute reduction methods also consider the uncertainties characterized by R_L, meaning that they have the same limitations for multi-label data as positive region reduct. Hence, it is necessary to reconsider attribute reduction methods for multi-label data to improve their ability to model the uncertainties implied in multi-label data.

4. New attribute reduction approach for multi-label data

This section introduces a new type of attribute reduct, referred to as *complementary decision reduct*. The relationships between complementary decision reduct and two representative types of attribute reducts will also be investigated, and significant advantages of complementary decision reduct in revealing the uncertainties implied in multi-label data will be demonstrated. Furthermore, a discernibility matrix-based method is introduced for computing all complementary decision reducts, and a heuristic algorithm is proposed for effectively computing a single complementary decision reduct.

4.1. Complementary decision reduct in multi-label decision table

To characterize the label information implied in multi-label data, the following definition is presented:

Definition 2. *Let $S = (U, A, L)$ be a multi-label decision table, where $A = \{a_1, a_2, \cdots, a_p\}$ and $L = \{l_1, l_2, \cdots, l_q\}$. Given a label $l_i \in L$, a label information set with respect to l_i is defined as follows:*

$$E_i = \{x \in U : \ l_i(x) = 1\}.$$

A label information set is the set of all objects having the label. According to Convention 2 regarding multi-label decision table, $\cup_{i=1}^{q} E_i = U$, i.e., $E_1, E_2, \cdots, E_q$ form a cover of U.

Two particular functions will now be presented to characterize the uncertainties implied in multi-label data.

Definition 3. *Let $S = (U, A, L)$ be a multi-label decision table, $P(L)$ be the power set of label set L, and $E_1, E_2, \cdots, E_q$ be q label information sets. Given a subset $B \subseteq A$, a coarse decision function $\mathcal{C}_B : U \to P(L)$ and a fine decision function $\mathcal{F}_B : U \to P(L)$ are defined as follows:*

$$\mathcal{C}_B(x) = \{l_i : x \in \overline{R_B}(E_i)\} = \{l_i : [x]_B \cap E_i \neq \emptyset\}, \quad x \in U,$$

$$\mathcal{F}_B(x) = \{l_i : x \in \underline{R_B}(E_i)\} = \{l_i : [x]_B \subseteq E_i\}, \quad x \in U.$$

A coarse decision function $\mathcal{C}_B(x)$ is the set of labels associated with at least one object in $[x]_B$. In other words, $\mathcal{C}_B(x)$ is the union of the label sets of all objects in $[x]_B$. Similarly, $\mathcal{F}_B(x)$ is the set of labels associated with all objects in $[x]_B$, i.e., the intersection of the label sets of all objects in $[x]_B$. Hence, coarse decision function and fine decision function represent, respectively, all possibly associated labels and all certainly associated labels for the objects in $[x]_B$. These functions are illustrated by the following example:

Example 3. *For the multi-label decision table $S = (U, A, L)$ given by Table 1, the coarse decision function and the fine decision function with respect to A can be calculated as follows:*

$$\begin{aligned}
&\mathcal{C}_A(x_1) = \mathcal{C}_A(x_3) = \{l_1, l_3\},\\
&\mathcal{C}_A(x_2) = \{l_2\},\\
&\mathcal{C}_A(x_4) = \mathcal{C}_A(x_5) = \mathcal{C}_A(x_7) = \{l_1, l_2, l_3\},
\end{aligned}$$

$$\begin{aligned}
&\mathcal{C}_A(x_6) = \mathcal{C}_A(x_8) = \{l_1, l_2, l_3\},\\
&\mathcal{C}_A(x_9) = \mathcal{C}_A(x_{11}) = \{l_1, l_2, l_3\},\\
&\mathcal{C}_A(x_{10}) = \{l_1, l_2, l_3\},
\end{aligned}$$

$$\begin{aligned}
&\mathcal{F}_A(x_1) = \mathcal{F}_A(x_3) = \{l_1\},\\
&\mathcal{F}_A(x_2) = \{l_2\},\\
&\mathcal{F}_A(x_4) = \mathcal{F}_A(x_5) = \mathcal{F}(x_7) = \{l_3\},\\
&\mathcal{F}_A(x_6) = \mathcal{F}_A(x_8) = \{l_2\},\\
&\mathcal{F}_A(x_9) = \mathcal{F}_A(x_{11}) = \{l_2\},\\
&\mathcal{F}_A(x_{10}) = \{l_1, l_2, l_3\}.
\end{aligned}$$

The following proposition shows some intuitive properties of coarse decision function and fine decision function:

Proposition 1. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B, C \subseteq A$. Then,*

(1) If $B \subseteq C$, then $\mathcal{F}_B(x) \subseteq \mathcal{F}_C(x) \subseteq \mathcal{C}_C(x) \subseteq \mathcal{C}_B(x)$.

(2) For any $x \in U$, $\mathcal{C}_B(x) \neq \emptyset$.

(3) If $[x]_B = [y]_B$, then $\mathcal{C}_B(x) = \mathcal{C}_B(y)$ and $\mathcal{F}_B(x) = \mathcal{F}_B(y)$.

Proof. (1) Because $B \subseteq C$, $[x]_C \subseteq [x]_B$.

If $l_i \in \mathcal{F}_B(x)$, then $x \in \underline{R_B}(E_i)$, i.e., $[x]_B \subseteq E_i$. Because $[x]_C \subseteq [x]_B$, $[x]_C \subseteq E_i$, and, hence, $l_i \in \mathcal{F}_C(x)$. Therefore, $\mathcal{F}_B(x) \subseteq \mathcal{F}_C(x)$.

If $l_j \in \mathcal{F}_C(x)$, then $x \in \underline{R_C}(E_j)$, i.e., $[x]_C \subseteq E_j$. It is possible to obtain $[x]_C \cap E_j \neq \emptyset$. Then, $l_j \in \mathcal{C}_C(x)$. Therefore, $\mathcal{F}_C(x) \subseteq \mathcal{C}_C(x)$.

If $l_k \in \mathcal{C}_C(x)$, then $x \in \overline{R_C}(E_k)$, i.e., $[x]_C \cap E_k \neq \emptyset$. Because $[x]_C \subseteq [x]_B$, it follows that $[x]_B \cap E_k \neq \emptyset$. Hence, $l_k \in \mathcal{C}_B(x)$. Therefore, $\mathcal{C}_C(x) \subseteq \mathcal{C}_B(x)$.

(2) If there exists $x \in U$ such that $\mathcal{C}_B(x) = \emptyset$, then $[x]_B \cap E_i = \emptyset$, $i = 1, 2, \cdots, q$. Hence, $[x]_B \cap (E_1 \cup E_2 \cup \cdots \cup E_q) = \emptyset$. On the other hand, because $\cup_{i=1}^{q} E_i = U$, it follows that $[x]_B \cap (E_1 \cup E_2 \cup \cdots \cup E_q) = [x]_B \cap U = [x]_B \neq \emptyset$. This is a contradiction.

(3) The desired conclusion is straightforward by the definitions of $\mathcal{C}_B(x)$ and $\mathcal{F}_B(x)$ and the fact that $[x]_B = [y]_B$.

□

Now, complementary decision consistent set can be defined using coarse decision function and fine decision function.

Definition 4. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. If $\mathcal{C}_B(x) = \mathcal{C}_A(x)$ and $\mathcal{F}_B(x) = \mathcal{F}_A(x)$ for all $x \in U$, it can be said that B is a complementary decision consistent set of S; otherwise, B is a complementary decision inconsistent set.*

According to Proposition 1 (1) and Definition 4, for a given multi-label decision table $S = (U, A, L)$ and $B \subseteq A$, it can be concluded that, if B is inconsistent, then any subset of B is also inconsistent.

Next, the definition of a complementary decision reduct is presented.

Definition 5. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. If B is a complementary decision consistent set and no proper subset of B is a complementary decision consistent set, then B is called a complementary decision reduct of S.*

A complementary decision reduct is the minimal set of condition attributes that simultaneously preserves the invariances of coarse decision function and fine decision function for all objects in U. In other words, a complementary decision reduct is an essential part of a multi-label decision table, which suffices to preserve the uncertainties implied in multi-label data. In general, a multi-label decision table may have more than one complementary decision reduct.

Intuitively, different condition attributes may play different roles in preserving the uncertainties implied in multi-label data. In the following discussion, these attributes are partitioned into two classes: dispensable attributes and indispensable attributes.

Definition 6. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. An attribute $a \in B$ is dispensable in B with respect to L if for all $x \in U$, $\mathcal{C}_B(x) = \mathcal{C}_{B-\{a\}}(x)$ and $\mathcal{F}_B(x) = \mathcal{F}_{B-\{a\}}(x)$; otherwise, a is indispensable in B with respect to L.*

If an attribute is dispensable in a multi-label decision table, it can be removed without changing the uncertainties of the decision table. However, an indispensable attribute carries information essential to the decision table and cannot be removed.

According to Definition 4 and Definition 6, it is easy to establish the relationship between indispensable attribute and inconsistent set, as shown in the following proposition:

Proposition 2. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$ be a complementary decision consistent set. Then, $a \in B$ is an indispensable attribute in B if and only if $B - \{a\}$ is an inconsistent set.*

Next, another fundamental concept of rough set theory, the core, which can be interpreted as the most characteristic part of the condition attributes in a multi-label decision table, will be presented.

Definition 7. *Let $S = (U, A, L)$ be a multi-label decision table. The set of all indispensable attributes in A is called the core of A and is denoted by $CORE(A)$.*

The following proposition establishes the relationship between the core and all complementary decision reducts:

Proposition 3. *Let $S = (U, A, L)$ be a multi-label decision table, and let $\{R_i | i = 1, \cdots, k\}$ be the set of all complementary decision reducts of S. Then, it follows that*

$$CORE(A) = \bigcap_{i=1}^{k} R_i.$$

Proof. Let $a \in CORE(A)$. Then, it can be concluded that $a \in R_i$ for every $i = 1, 2, \cdots, k$. Otherwise, there exists one reduct R_i such that $a \notin R_i$. Because $R_i = R_i - \{a\}$, $R_i - \{a\}$ is also a reduct. This implies that $R_i - \{a\}$ is a consistent set. On the other hand, $a \in CORE(A)$ means that a is an indispensable attribute in A. Note that A is also a consistent set; hence, by Proposition 2, $A - \{a\}$ is an inconsistent set. Therefore, $R_i - \{a\} \subseteq A - \{a\}$ is an inconsistent set. This is a contradiction.

Conversely, let $a \in \bigcap_{i=1}^{k} R_i$. Suppose that $a \notin CORE(A)$, i.e., a is a dispensable attribute in A. Because A is a consistent set, $A - \{a\}$ is a complementary decision consistent set of S. In this case, there must exist at least one reduct R such that $R \subseteq A - \{a\}$. Therefore, $a \in \bigcap_{i=1}^{k} R_i \subseteq R \subseteq A - \{a\}$, i.e., $a \in A - \{a\}$, which is a contradiction. □

Proposition 3 states that the core is contained in every complementary decision reduct. Figure 1 provides a visual presentation of the relationship between the core and all complementary decision reducts.

Now, the relationship between complementary decision reduct and positive region reduct will be investigated.

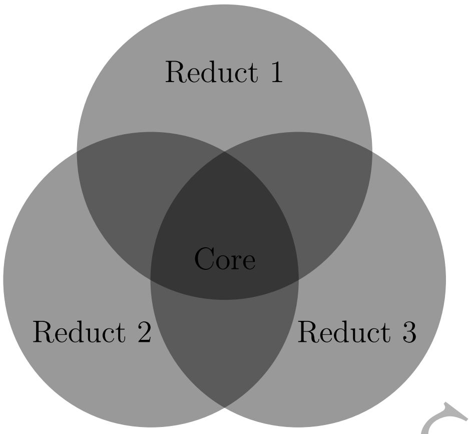

Figure 1: Relationship between the core and all complementary decision reducts

Theorem 1. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. If B is a complementary decision reduct of S, then B must be a positive region reduct of S.*

Proof. For any $A^{'} \subseteq A$, according to the definition of positive region, it is known that $x \in POS_{A'}(D)$ is equivalent to $[x]_{A'} \subseteq POS_{A'}(D)$, and, moreover, that $[x]_{A'} \subseteq POS_{A'}(D)$ if and only if the label sets of all objects in $[x]_{A'}$ are identical. Combining the definitions of $\mathcal{C}_{A'}(x)$ and $\mathcal{F}_{A'}(x)$, it follows that $x \in POS_{A'}(D)$ if and only if $\mathcal{C}_{A'}(x) = \mathcal{F}_{A'}(x)$. This means that if $x \in POS_{A'}(D)$, then the label set of every object in $[x]_{A'}$ must be $\mathcal{C}_{A'}(x)$.

Let $B \subseteq A$ be a complementary decision reduct of S. Then, for all $x \in U$, the following assertions can be made:

(1) $\mathcal{C}_A(x) = \mathcal{C}_B(x)$ and $\mathcal{F}_A(x) = \mathcal{F}_B(x)$.

(2) For any $B^{'} \subset B$, $B^{'}$ is a complementary decision inconsistent set.

By Assertion (1),

$$\begin{aligned} x \in POS_A(D) &\iff \mathcal{C}_A(x) = \mathcal{F}_A(x) \\ &\iff \mathcal{C}_B(x) = \mathcal{C}_A(x) = \mathcal{F}_A(x) = \mathcal{F}_B(x) \\ &\iff x \in POS_B(D). \end{aligned}$$

That is, $POS_A(D) = POS_B(D)$. This implies that B is also a positive region consistent set.

Next, it is proven that B must also be a positive region reduct. If not, suppose that there exists $B^{'} \subset B$ such that $POS_A(D) = POS_{B'}(D)$. Let

$x \in POS_A(D)$. Then, it follows that $\mathcal{C}_A(x) = \mathcal{F}_A(x)$, and the label set of every object in $[x]_A$ is $\mathcal{C}_A(x)$. Because $POS_A(D) = POS_{B'}(D)$, it is known that $\mathcal{C}_{B'}(x) = \mathcal{F}_{B'}(x)$ and that the label set of every object in $[x]_{B'}$ is $\mathcal{C}_{B'}(x)$. Considering $B^{'} \subset B \subseteq A$, it follows that $[x]_A \subseteq [x]_{B'}$. Hence, the label set of every object in $[x]_A$ is $\mathcal{C}_{B'}(x)$. This means that $\mathcal{C}_A(x) = \mathcal{C}_{B'}(x)$. Therefore, $\mathcal{C}_A(x) = \mathcal{C}_{B'}(x) = \mathcal{F}_{B'}(x) = \mathcal{F}_A(x)$, which contradicts Assertion (2). Therefore, B is a positive region reduct. □

Note that the converse of Theorem 1 is not always true, i.e., a positive region reduct may not be a complementary decision reduct, as illustrated by Example 4.

Example 4 (Continued from Example 3)**.** *For the multi-label decision table $S = (U, A, L)$ given by Table 1,*

$$\begin{aligned}
&\mathcal{C}_{\{a\}}(x_1) = \{l_1, l_2, l_3\} \neq \mathcal{C}_A(x_1),\\
&\mathcal{C}_{\{b\}}(x_2) = \{l_1, l_2, l_3\} \neq \mathcal{C}_A(x_2),\\
&\mathcal{C}_{\{c\}}(x_2) = \{l_1, l_2, l_3\} \neq \mathcal{C}_A(x_2),\\
&\mathcal{F}_{\{a,b\}}(x_1) = \emptyset \neq \mathcal{F}_A(x_1),\\
&\mathcal{C}_{\{b,c\}}(x_2) = \{l_1, l_2, l_3\} \neq \mathcal{C}_A(x_2),\\
&\mathcal{C}_{\{a,c\}}(x) = \mathcal{C}_A(x),\\
&\mathcal{F}_{\{a,c\}}(x) = \mathcal{F}_A(x) \ for\ any\ x \in U.
\end{aligned}$$

Therefore, the unique complementary decision reduct $\{a, c\}$ can be obtained.

Considering Example 2, it is known that $\{a, b\}$ and $\{a, c\}$ are two positive region reducts. Therefore, the positive region reduct $\{a, c\}$ is a complementary decision reduct, whereas $\{a, b\}$ is not.

Furthermore, it is known that $\{a, c\}$ is more valuable in reducing unnecessary attributes than $\{a, b\}$. Hence, the complementary decision reduct is more appropriate for multi-label data than the positive region reduct. The reason for this is that the coarse decision function and the fine decision function represent, respectively, all possibly associated labels and all certainly associated labels for each object in U and, hence, can more reasonably characterize the uncertainties implied in multi-label data than the indiscernibility relation R_L.

Note that most existing attribute reduction methods also consider the uncertainties characterized by R_L. Therefore, for multi-label data, complementary decision reduct has significant advantages over most existing attribute reduction methods.

4.2. Relationship between complementary decision reduct and generalized decision reduct

Complementary decision reduct may be easily confused with generalized decision reduct [21, 40], which is one of the important types of attribute reducts in rough set theory.

For a multi-label decision table $S = (U, A, L)$, each label attribute can be viewed as a binary decision attribute, and then it can be considered as a special type of decision table. The concept of generalized decision reduct can be rewritten as follows:

Let $S = (U, A, L)$ be a multi-label decision table, where $A = \{a_1, a_2, \cdots, a_p\}$ and $L = \{l_1, l_2, \cdots, l_q\}$. Given a subset $B \subseteq A$, a generalized decision function with respect to B is defined by $\partial_B : U \to P(\times_{l \in L} V_l)$ with

$$\partial_B(x) = \{L(y) : y \in [x]_B\}, \;\; x \in U,$$

where $P(\times_{l \in L} V_l)$ is the powerset of Cartesian product $\times_{l \in L} V_l$ and $L(y) = (l_1(y), l_2(y), \cdots, l_q(y))$ is a Boolean value vector, which represents the set of labels associated with y.

Let $S = (U, A, L)$ be a multi-label decision table. A subset $B \subseteq A$ is a generalized decision consistent set of S if and only if $\partial_B(x) = \partial_A(x)$ for all $x \in U$. If B is a minimal generalized decision consistent set (with respect to inclusion), then B is called a generalized decision reduct of S.

The following example can be used to illustrate the difference and the relationship between complementary decision reduct and generalized decision reduct:

Example 5. *For the multi-label decision table $S = (U, A, L)$ given in Table 1, it is known that*

$$\begin{aligned}
U/R_A &= \{X_1, X_2, X_3, X_4, X_5, X_6\} \\
&= \{\{x_1, x_3\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_6, x_8\}, \{x_9, x_{11}\}, \{x_{10}\}\}, \\
U/R_{\{a,b\}} &= \{X_1 \cup X_4, X_2, X_3, X_5, X_6\} \\
&= \{\{x_1, x_3, x_6, x_8\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_9, x_{11}\}, \{x_{10}\}\}, \\
U/R_{\{a,c\}} &= \{X_1, X_2, X_3, X_4 \cup X_5, X_6\} \\
&= \{\{x_1, x_3\}, \{x_2\}, \{x_4, x_5, x_7\}, \{x_6, x_8, x_9, x_{11}\}, \{x_{10}\}\}, \\
U/R_{\{b,c\}} &= \{X_1, X_2 \cup X_4, X_3, X_5, X_6\} \\
&= \{\{x_1, x_3\}, \{x_2, x_6, x_8\}, \{x_4, x_5, x_7\}, \{, x_9, x_{11}\}, \{x_{10}\}\}.
\end{aligned}$$

Hence, the generalized decision function with respect to A can be calculated as follows:

$$\begin{aligned}\partial_A(x_1) &= \partial_A(x_3) = \{(1,0,0),(1,0,1)\}\\ &= \{\{l_1\},\{l_1,l_3\}\},\\ \partial_A(x_2) &= \{(0,1,0)\} = \{\{l_2\}\},\\ \partial_A(x_4) &= \partial_A(x_5) = \partial_A(x_7) = \{(1,0,1),(0,0,1),(1,1,1)\}\\ &= \{\{l_1,l_3\},\{l_3\},\{l_1,l_2,l_3\}\},\\ \partial_A(x_6) &= \partial_A(x_8) = \{(0,1,0),(1,1,1)\}\\ &= \{\{l_2\},\{l_1,l_2,l_3\}\},\\ \partial_A(x_9) &= \partial_A(x_{11}) = \{(0,1,1),(1,1,0)\}\\ &= \{\{l_2,l_3\},\{l_1,l_2\}\},\\ \partial_A(x_{10}) &= \{(1,1,1)\} = \{\{l_1,l_2,l_3\}\}.\end{aligned}$$

Furthermore, it can be determined that

$$\begin{aligned}\partial_{\{a,b\}}(x_6) &= \{(1,0,0),(1,0,1),(0,1,0),(1,1,1)\}\\ &= \{\{l_1\},\{l_1,l_3\},\{l_2\},\{l_1,l_2,l_3\}\}\\ &\neq \partial_A(x_6),\\ \partial_{\{a,c\}}(x_6) &= \{(0,1,0),(1,1,1),(0,1,1),(1,1,0)\}\\ &= \{\{l_2\},\{l_1,l_2,l_3\},\{l_2,l_3\},\{l_1,l_2\}\}\\ &\neq \partial_A(x_6),\\ \partial_{\{b,c\}}(x_2) &= \{(0,1,0),(1,1,1)\}\\ &= \{\{l_2\},\{l_1,l_2,l_3\}\}\\ &\neq \partial_A(x_2).\end{aligned}$$

It means that each of the attribute subsets $\{a,b\}$, $\{a,c\}$, and $\{b,c\}$ is not a generalized decision reduct of S. In other words, no attribute can be removed from A under the condition $\partial_B(x) = \partial_A(x)$ for all $x \in U$. Consequently, there exists the unique generalized decision reduct $\{a,b,c\}$.

Now, let us compute a complementary decision reduct of S. Comparing U/R_A with $U/R_{\{a,c\}}$, it is easy to know that we need to check the condition of complementary decision reduct only for X_4 and X_5. The unions of the label sets of $X_4 = \{x_6,x_8\}$ and $X_5 = \{x_9,x_{11}\}$ are equal, i.e., $\{l_1,l_2,l_3\}$, and the intersections of the label sets of them are also equal, i.e., $\{l_2\}$. Thus X_4 and

X_5 can be safely merged without altering the unions and the intersections of all label sets of the equivalence classes X_4 and X_5 in U/R_A. In other words, the attribute b can be removed from A. The further examination shows that $\{a, c\}$ is also minimal. As a result, $\{a, c\}$ is a complementary decision reduct of S.

In fact, more mathematically viewing the definitions of complementary decision reduct and generalized decision reduct can bring a profound comprehension about the essential difference of them.

Let $V = (v_1, v_2, \cdots, v_q)$ and $W = (w_1, w_2, \cdots, w_q)$ be two q-dimensional Boolean value vectors. Define the disjunction (and) and the conjunction (or) of V and W as follows:

$$V \vee W = (v_1 \vee w_1, v_2 \vee w_2, \cdots, v_q \vee w_q),$$

$$V \wedge W = (v_1 \wedge w_1, v_2 \wedge w_2, \cdots, v_q \wedge w_q).$$

Consider the definition conditions of $\mathcal{C}_B(x)$ and $\mathcal{F}_B(x)$ in Definition 3. It should be note that $[x]_B \cap E_i \neq \emptyset$ if and only if there exists $y \in [x]_B$ such that $l_i(y) = 1$. So $\vee\partial_B(x) = \vee\{L(y) : y \in [x]_B\}$ represents $\mathcal{C}_B(x)$ in semantical equivalence; in other words, both $\vee\partial_B(x)$ and $\mathcal{C}_B(x)$ represent the set of labels associated with at least one object in $[x]_B$, i.e., the union of the label sets of all objects in $[x]_B$. Analogously, $[x]_B \subseteq E_i$ if and only if $l_i(y) = 1$ for any $y \in [x]_B$. So $\wedge\partial_B(x) = \wedge\{L(y) : y \in [x]_B\}$ represents $\mathcal{F}_B(x)$ in semantical equivalence; in other words, both $\wedge\partial_B(x)$ and $\mathcal{F}_B(x)$ represent the set of labels associated with all objects in $[x]_B$, i.e., the intersection of the label sets of all objects in $[x]_B$.

For a multi-label decision table $S = (U, A, L)$, the above analysis indicates that a generalized decision reduct B is requested to preserve the label vector collection $\partial_A(x) = \{L(y) : y \in [x]_A\}$, i.e., $\partial_B(x) = \partial_A(x)$ for all $x \in U$. Contrastively, a complementary decision reduct B is requested to simultaneously preserve the results of $\partial_A(x)$ under the operations $\vee$ and $\wedge$, $\mathcal{C}_A(x)$ and $\mathcal{F}_A(x)$, i.e., $\mathcal{C}_B(x) = \mathcal{C}_A(x)$ and $\mathcal{F}_B(x) = \mathcal{F}_A(x)$ for all $x \in U$.

Note that preserving the label vector collection $\partial_A(x) = \{L(y) : y \in [x]_A\}$ apparently implies preserving the two label subsets $\mathcal{C}_A(x)$ and $\mathcal{F}_A(x)$, so a generalized decision consistent set must be a complementary decision consistent set, but the converse is always not true, i.e., a complementary decision consistent set may not be a generalized decision consistent set. This means that a generalized decision reduct must be a complementary decision consistent set, but not necessarily be a complementary decision reduct; a complementary decision reduct is even not a generalized decision consistent set. Moreover, because complementary decision reduct searches a minimal attribute subset in a larger range than that in which generalized decision reduct does, complementary decision reduct may be more compact than generalized decision reduct.

4.3. Discernibility matrix of complementary decision reduct

This section presents a discernibility matrix-based method [38] for computing all complementary decision reducts. For purposes of discussion, two lemmas on coarse decision function and fine decision function are presented.

Lemma 1. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Then, the following conditions are equivalent:*

(1) For any $x \in U$, $\mathcal{C}_B(x) = \mathcal{C}_A(x)$.

(2) For any $x, y \in U$, if $\mathcal{C}_A(x) \neq \mathcal{C}_A(y)$, then $[x]_B \cap [y]_B = \emptyset$.

Proof. "(1) $\Longrightarrow$ (2)". If there exist $x, y \in U$ such that $[x]_B \cap [y]_B \neq \emptyset$, then $[x]_B = [y]_B$. By Proposition 1 (3), it follows that $\mathcal{C}_B(x) = \mathcal{C}_B(y)$. Note that $\mathcal{C}_B(x) = \mathcal{C}_A(x)$ and that $\mathcal{C}_B(y) = \mathcal{C}_A(y)$. Then, it follows that $\mathcal{C}_A(x) = \mathcal{C}_A(y)$.

"(2) $\Longrightarrow$ (1)". By Proposition 1 (1), $\mathcal{C}_A(x) \subseteq \mathcal{C}_B(x)$ for any $x \in U$. Hence, it is necessary only to prove $\mathcal{C}_B(x) \subseteq \mathcal{C}_A(x)$ for any $x \in U$.

If $l_i \in \mathcal{C}_B(x)$, then $[x]_B \cap E_i \neq \emptyset$. Suppose that $y \in [x]_B \cap E_i$. Because $y \in [x]_B$, $[y]_B \cap [x]_B \neq \emptyset$. According to Condition (2), $\mathcal{C}_A(x) = \mathcal{C}_A(y)$. Moreover, because $y \in E_i$, it follows that $[y]_A \cap E_i \neq \emptyset$. This means that $l_i \in \mathcal{C}_A(y) = \mathcal{C}_A(x)$. Therefore, it can be concluded that $\mathcal{C}_B(x) \subseteq \mathcal{C}_A(x)$ for any $x \in U$.

Hence, $\mathcal{C}_B(x) = \mathcal{C}_A(x)$ holds for any $x \in U$. □

Similarly to the proof of Lemma 1, the following lemma can be derived:

Lemma 2. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Then, the following conditions are equivalent:*

(1) For any $x \in U$, $\mathcal{F}_B(x) = \mathcal{F}_A(x)$.

(2) For any $x, y \in U$, if $\mathcal{F}_A(x) \neq \mathcal{F}_A(y)$, then $[x]_B \cap [y]_B = \emptyset$.

Now, from Lemmas 1 and 2, it is possible to derive the following judgment theorem regarding complementary decision consistent set:

Theorem 2 (Judgment theorem of complementary decision consistent set)**.** *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Then, B is a complementary decision consistent set if and only if, for any $x, y \in U$, the following two conditions both hold:*

(1) If $\mathcal{C}_A(x) \neq \mathcal{C}_A(y)$, then $[x]_B \cap [y]_B = \emptyset$.

(2) If $\mathcal{F}_A(x) \neq \mathcal{F}_A(y)$, then $[x]_B \cap [y]_B = \emptyset$.

Theorem 2 also facilitates a discernibility matrix-based method for computing all complementary decision reducts.

Definition 8. *Let $S = (U, A, L)$ be a multi-label decision table, and let $U/R_A = \{X_1, X_2, \cdots, X_m\}$. Denote*

$$\Delta = \{([x]_A, [y]_A) : \quad \mathcal{C}_A(x) \neq \mathcal{C}_A(y) \text{ or } \mathcal{F}_A(x) \neq \mathcal{F}_A(y)\}.$$

Denote by $a_k(X_i)$ the value of a_k with respect to the objects in X_i, and define

$$M(X_i, X_j) = \begin{cases} \{a_k \in A : a_k(X_i) \neq a_k(X_j)\}, & (X_i, X_j) \in \Delta; \\ A, & (X_i, X_j) \notin \Delta. \end{cases}$$

Then, $M(X_i, X_j)$ is called a complementary decision discernibility attribute set and $\mathbb{M} = (M(X_i, X_j), i, j \leq m)$ is called a complementary decision discernibility matrix.

For the complementary decision discernibility matrix, the following property can be established:

Proposition 4. *A discernibility matrix $\mathbb{M} = (M(X_i, X_j), i, j \leq m)$ satisfies the following properties:*

(1) $\mathbb{M}$ is a symmetric matrix, i.e., for any $i, j \leq m$, $M(X_i, X_j) = M(X_j, X_i)$.

(2) Elements on the main diagonals are all A, i.e., for any $i \leq m$, $M(X_i, X_i) = A$.

(3) For any $i, s, j \leq m$, $M(X_i, X_j) \subseteq M(X_i, X_s) \cup M(X_s, X_j)$.

Proof. The proofs of (1) and (2) are straightforward. It is necessary only to prove (3). Suppose that there exists $a_k \in A$ such that if $a_k \in M(X_i, X_j)$, then $a_k \notin M(X_i, X_s) \cup M(X_s, X_j)$. According to Definition 8, it follows that $a_k(X_i) = a_k(X_s)$ and $a_k(X_s) = a_k(X_j)$. Hence, $a_k(X_i) = a_k(X_j)$, i.e., $a_k \notin M(X_i, X_j)$, which is a contradiction. □

The following discussion establishes the connection between complementary decision consistent set and discernibility matrix:

Proposition 5. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Then, B is a complementary decision consistent set if and only if $B \cap M(X_i, X_j) \neq \emptyset$ for all $(X_i, X_j) \in \Delta$.*

Proof. "$\Longrightarrow$". For any $(X_i, X_j) \in \Delta$, there exist $x, y \in U$ such that $X_i = [x]_A$ and $X_j = [y]_A$. From the definition of Δ, it follows that $\mathcal{C}_A(x) \neq \mathcal{C}_A(y)$ or $\mathcal{F}_A(x) \neq \mathcal{F}_A(y)$. Because B is a complementary decision consistent set, it can be concluded that $[x]_B \cap [y]_B = \emptyset$ by Theorem 2. Therefore, there exists $a_k \in B$ such that $a_k(x) \neq a_k(y)$, i.e., $a_k(X_i) \neq a_k(X_j)$. Hence, $a_k \in M(X_i, X_j)$, i.e., $B \cap M(X_i, X_j) \neq \emptyset$.

"$\Longleftarrow$". Let $(X_i, X_j) \in \Delta$. Because $B \cap M(X_i, X_j) \neq \emptyset$ holds for all $(X_i, X_j) \in \Delta$, there exists $a_l \in B$ such that $a_l \in M(X_i, X_j)$. Then, it follows that $a_l(X_i) \neq a_l(X_j)$, i.e., $a_l(x) \neq a_l(y)$ for $[x]_A = X_i$ and $[y]_A = X_j$. This fact yields $[x]_B \cap [y]_B = \emptyset$. It can then be concluded that, if $(X_i, X_j) \in \Delta$, i.e., $\mathcal{C}_A(x) \neq \mathcal{C}_A(y)$ or $\mathcal{F}_A(x) \neq \mathcal{F}_A(y)$, then $[x]_B \cap [y]_B = \emptyset$. It then follows from Theorem 2 that B is a complementary decision consistent set. □

Next, the concept of a discernibility function for computing complementary decision reduct is introduced.

Definition 9. *Let $S = (U, A, L)$ be a multi-label decision table, and let $\mathbb{M} = (M(X_i, X_j), i, j \leq m)$ be a complementary decision discernibility matrix, where $A = \{a_1, a_2, \cdots, a_p\}$. A complementary decision discernibility function F_S for a multi-label decision table S is a Boolean function of p Boolean variables $\tilde{a}_1, \cdots, \tilde{a}_p$ corresponding to the attributes $a_1, \cdots, a_p$, respectively, and is defined as follows:*

$$\begin{aligned} F_S(\tilde{a}_1, \cdots, \tilde{a}_p) = & \bigwedge\{\bigvee M(X_i, X_j)\ i, j \leq m\} \\ = & \bigwedge\{\bigvee M(X_i, X_j),\ (X_i, X_j) \in \Delta\}, \end{aligned}$$

where $\bigvee M(X_i, X_j)$ is the disjunction of all variables $\tilde{a}$ such that $a \in M(X_i, X_j)$.

In the following discussion, a_i will be written instead of $\tilde{a}_i$ when no confusion arises.

The complementary decision discernibility function can be used to discern the set of reducts, as shown by the following proposition:

Proposition 6. *Let $S = (U, A, L)$ be a multi-label decision table. Then, a subset $B \subseteq A$ is a complementary decision reduct of S if and only if $\wedge B$ is a prime implicant of F_S.*

Proposition 6 provides a discernibility matrix-based method for computing all complementary decision reducts. The following example illustrates the validity of this approach:

Example 6. *Consider the multi-label decision table given by Table 1. It follows that $U/R_A = \{X_1, X_2, \cdots, X_6\}$, where*

$$X_1 = \{x_1, x_3\}, \quad X_2 = \{x_2\}, \quad X_3 = \{x_4, x_5, x_7\},$$
$$X_4 = \{x_6, x_8\},\ X_5 = \{x_9, x_{11}\},\ X_6 = \{x_{10}\}.$$

According to the calculated results of $\mathcal{C}_A(x)$ and $\mathcal{F}_A(x)$ in Example 3,

$$\begin{aligned}\Delta \ = \ & \{(X_1, X_2), (X_1, X_3), (X_1, X_4), (X_1, X_5), (X_1, X_6), (X_2, X_3), (X_2, X_4),\\ & (X_2, X_5), (X_2, X_6), (X_3, X_4), (X_3, X_5), (X_3, X_6), (X_4, X_6), (X_5, X_6)\}.\end{aligned}$$

Note that $\mathcal{C}_A(x_6) = \mathcal{C}_A(x_8) = \mathcal{C}_A(x_9) = \mathcal{C}_A(x_{11})$ and $\mathcal{F}_A(x_6) = \mathcal{F}_A(x_8) = \mathcal{F}_A(x_9) = \mathcal{F}_A(x_{11})$. Therefore, $(X_4, X_5) \notin \Delta$.

The complementary decision discernibility matrix can then be calculated, with the results shown in Table 2.

Table 2: Complementary decision discernibility matrix $\mathbb{M}$

	X_1	X_2	X_3	X_4	X_5	X_6
X_1						
X_2	a, c					
X_3	a, b	a, b, c				
X_4	c	a	a, b, c			
X_5	b, c	a, b	a, b, c	a, b, c		
X_6	a, b	b, c	a, b	a, b, c	a, c	

Consequently,

$$\begin{aligned} F_S = \ & (a \vee b \vee c) \wedge (a \vee c) \wedge (a \vee b) \wedge (c) \wedge (b \vee c) \wedge (a)\\ = \ & a \wedge c.\end{aligned}$$

By Proposition 6, it is known that $\{a, c\}$ is the unique complementary decision reduct, which is in accordance with the result in Example 4.

4.4. Heuristic algorithm for computing complementary decision reduct

According to the discussion in Section 4.3, it is known that all complementary decision reducts can be obtained from the prime implicants of the discernibility function. However, finding all the reducts or an optimal reduct (i.e., a reduct with the minimum number of attributes) has been proved to be an NP-hard problem [46]. This section discusses how to obtain a single heuristic "optimal" reduct using certain heuristic techniques.

First, a dependency function is introduced to characterize the degree of dependency of an attribute subset with respect to label set L in a given multi-label decision table.

Definition 10. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Denote*

$$\phi(B) = \begin{cases} \lambda, & \text{if } \mathcal{F}_B(x) = \emptyset \text{ for any } x \in U; \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda \in (0, 1)$ is a constant. The dependency function of L with respect to B is defined by [1]

$$\gamma_L(B) = \frac{\sum_{x \in U} |\mathcal{F}_B(x)| + \phi(B)}{\sum_{x \in U} |\mathcal{C}_B(x)|}.$$

The dependency function $\gamma_L(B)$ reflects the ability of B to approximate L or the degree of dependency of L on B. If for any $x \in U$, $\mathcal{F}_B(x) = \emptyset$ and $\mathcal{C}_B(x) = L$, then $\gamma_L(B)$ reaches the minimum value $\frac{\lambda}{|L| \cdot |U|}$. If for any $x \in U$, $\mathcal{F}_B(x) = \mathcal{C}_B(x)$, then $\gamma_L(B)$ reaches the maximum value of 1. Hence, $\frac{\lambda}{|L| \cdot |U|} \leq \gamma_L(B) \leq 1$.

If $\gamma_L(B) = 1$, i.e., for any $x \in U$, $\mathcal{F}_B(x) = \mathcal{C}_B(x)$, then the label sets of all objects in $[x]_B$ are certain. It can then be said that L totally depends on B, which is denoted by $B \Longrightarrow L$; if $\frac{\lambda}{|L| \cdot |U|} \leq \gamma_L(B) < 1$, it can be said that L partially depends on B, which is denoted by $B \Longrightarrow^{\gamma} L$.

For the dependency function, the following property can be established:

Proposition 7. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B_1, B_2 \subseteq A$. If $B_1 \subseteq B_2$, then $\gamma_L(B_1) \leq \gamma_L(B_2)$.*

[1] Note that $\sum_{x \in U} |\mathcal{C}_B(x)| > 0$ according to Proposition 1 (2).

Proof. First, assume that there exists $x \in U$ such that $\mathcal{F}_{B_1}(x) \neq \emptyset$. Noting that $B_1 \subseteq B_2$, by Proposition 1 (1), it follows that $\emptyset \neq \mathcal{F}_{B_1}(x) \subseteq \mathcal{F}_{B_2}(x) \subseteq \mathcal{C}_{B_2}(x) \subseteq \mathcal{C}_{B_1}(x)$. This means that $0 < \sum_{x\in U} |\mathcal{F}_{B_1}(x)| \leq \sum_{x\in U} |\mathcal{F}_{B_2}(x)| \leq \sum_{x\in U} |\mathcal{C}_{B_2}(x)| \leq \sum_{x\in U} |\mathcal{C}_{B_1}(x)|$. Therefore,

$$\gamma_L(B_1) = \frac{\sum_{x\in U} |\mathcal{F}_{B_1}(x)|}{\sum_{x\in U} |\mathcal{C}_{B_1}(x)|} \leq \frac{\sum_{x\in U} |\mathcal{F}_{B_2}(x)|}{\sum_{x\in U} |\mathcal{C}_{B_2}(x)|} = \gamma_L(B_2).$$

Next, assume that $\mathcal{F}_{B_1}(x) = \emptyset$ for any $x \in U$. Then,

$$\gamma_L(B_1) = \frac{\lambda}{\sum_{x\in U} |\mathcal{C}_{B_1}(x)|} \leq \frac{\sum_{x\in U} |\mathcal{F}_{B_2}(x)| + \phi(B_2)}{\sum_{x\in U} |\mathcal{C}_{B_2}(x)|} = \gamma_L(B_2).$$

Therefore, it follows that $\gamma_L(B_1) \leq \gamma_L(B_2)$. □

Proposition 7 states that dependency function increases monotonically with the number of attributes. In other words, as the number of attributes increases, the ability of condition attributes to approximate label set L also monotonically increases.

The following theorem states that complementary decision consistent set can be fully characterized by dependency function:

Theorem 3. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. Then, B is a complementary decision consistent set if and only if $\gamma_L(A) = \gamma_L(B)$.*

Proof. Note that $B \subseteq A$. By Proposition 1 (1), it follows that $\mathcal{F}_B(x) \subseteq \mathcal{F}_A(x) \subseteq \mathcal{C}_A(x) \subseteq \mathcal{C}_B(x)$ for any $x \in U$.

"$\Longrightarrow$". Suppose that B is a complementary decision consistent set. Then, for any $x \in U$, it follows that $\mathcal{C}_A(x) = \mathcal{C}_B(x)$ and $\mathcal{F}_A(x) = \mathcal{F}_B(x)$. This implies that $\phi(A) = \phi(B)$. Therefore,

$$\gamma_L(A) = \frac{\sum_{x\in U} |\mathcal{F}_A(x)| + \phi(A)}{\sum_{x\in U} |\mathcal{C}_A(x)|} = \frac{\sum_{x\in U} |\mathcal{F}_B(x)| + \phi(B)}{\sum_{x\in U} |\mathcal{C}_B(x)|} = \gamma_L(B).$$

"$\Longleftarrow$". Consider the following two cases:

Case 1. There exists $x \in U$ such that $\mathcal{F}_B(x) \neq \emptyset$.

Note that $\mathcal{F}_B(x) \subseteq \mathcal{F}_A(x)$. It follows that $\mathcal{F}_A(x) \neq \emptyset$. Recall that $\mathcal{F}_B(x) \subseteq \mathcal{F}_A(x) \subseteq \mathcal{C}_A(x) \subseteq \mathcal{C}_B(x)$ for any $x \in U$; it follows that

$$0 < \sum_{x\in U} |\mathcal{F}_B(x)| \leq \sum_{x\in U} |\mathcal{F}_A(x)| \leq \sum_{x\in U} |\mathcal{C}_A(x)| \leq \sum_{x\in U} |\mathcal{C}_B(x)|.$$

Therefore,

$$\gamma_L(B) = \frac{\sum_{x\in U} |\mathcal{F}_B(x)|}{\sum_{x\in U} |\mathcal{C}_B(x)|} \leq \frac{\sum_{x\in U} |\mathcal{F}_A(x)|}{\sum_{x\in U} |\mathcal{C}_B(x)|} \leq \frac{\sum_{x\in U} |\mathcal{F}_A(x)|}{\sum_{x\in U} |\mathcal{C}_A(x)|} = \gamma_L(A).$$

Combining this result with $\gamma_L(A) = \gamma_L(B)$, it follows that

$$\frac{\sum_{x\in U} |\mathcal{F}_B(x)|}{\sum_{x\in U} |\mathcal{C}_B(x)|} = \frac{\sum_{x\in U} |\mathcal{F}_A(x)|}{\sum_{x\in U} |\mathcal{C}_B(x)|} = \frac{\sum_{x\in U} |\mathcal{F}_A(x)|}{\sum_{x\in U} |\mathcal{C}_A(x)|}.$$

This means that

$$\sum_{x\in U} |\mathcal{F}_A(x)| = \sum_{x\in U} |\mathcal{F}_B(x)| \;\text{ and }\; \sum_{\mathrm{x}\in\mathrm{U}} |\mathcal{C}_{\mathrm{A}}(\mathrm{x})| = \sum_{\mathrm{x}\in\mathrm{U}} |\mathcal{C}_{\mathrm{B}}(\mathrm{x})|.$$

It is claimed here that, for any $x \in U$, $\mathcal{F}_A(x) = \mathcal{F}_B(x)$ holds. In fact, if there exists some $x^* \in U$ such that $\mathcal{F}_B(x^*) \subset \mathcal{F}_A(x^*)$, then $|\mathcal{F}_B(x^*)| < |\mathcal{F}_A(x^*)|$. Note that $\mathcal{F}_B(x) \subseteq \mathcal{F}_A(x)$ holds for any $x \in U$; it follows that $\sum_{x\in U} |\mathcal{F}_B(x)| < \sum_{x\in U} |\mathcal{F}_A(x)|$, which is a contradiction. Similarly, it can be proved that $\mathcal{C}_A(x) = \mathcal{C}_B(x)$ for any $x \in U$. Hence, B is a complementary decision consistent set.

Case 2. $\mathcal{F}_B(x) = \emptyset$ for any $x \in U$.

Note that $\mathcal{C}_A(x) \subseteq \mathcal{C}_B(x)$ holds for any $x \in U$. Combining this with Proposition 1 (2), it follows that $0 < \sum_{x\in U} |\mathcal{C}_A(x)| \leq \sum_{x\in U} |\mathcal{C}_B(x)|$. Hence,

$$\frac{1}{\sum_{x\in U} |\mathcal{C}_A(x)|} \geq \frac{1}{\sum_{x\in U} |\mathcal{C}_B(x)|}.$$

Note that $\sum_{x\in U} |\mathcal{F}_A(x)| + \phi(A) \geq \lambda$; it follows that

$$\gamma_L(B) = \frac{\lambda}{\sum_{x\in U} |\mathcal{C}_B(x)|} \leq \frac{\lambda}{\sum_{x\in U} |\mathcal{C}_A(x)|} \leq \frac{\sum_{x\in U} |\mathcal{F}_A(x)| + \phi(A)}{\sum_{x\in U} |\mathcal{C}_A(x)|} = \gamma_L(A).$$

Combining this result with $\gamma_L(A) = \gamma_L(B)$, it follows that

$$\frac{\lambda}{\sum_{x\in U} |\mathcal{C}_B(x)|} = \frac{\lambda}{\sum_{x\in U} |\mathcal{C}_A(x)|} = \frac{\sum_{x\in U} |\mathcal{F}_A(x)| + \phi(A)}{\sum_{x\in U} |\mathcal{C}_A(x)|}.$$

This implies that

$$\sum_{x\in U} |\mathcal{C}_B(x)| = \sum_{x\in U} |\mathcal{C}_A(x)|, \sum_{x\in U} |\mathcal{F}_A(x)| + \phi(A) = \lambda < 1.$$

Because $\sum_{x\in U}|\mathcal{F}_A(x)|$ is a nonnegative integer and $\phi(A)\geq 0$, it follows that $\sum_{x\in U}|\mathcal{F}_A(x)|=0$. This means that $\mathcal{F}_A(x)=\emptyset$ holds for any $x\in U$, i.e., that $\mathcal{F}_A(x)=\mathcal{F}_B(x)=\emptyset$ for any $x\in U$. In addition, it is claimed here that, for any $x\in U$, $\mathcal{C}_A(x)=\mathcal{C}_B(x)$ holds. In fact, if there exists $x^{**}\in U$ such that $\mathcal{C}_A(x^{**})\subset\mathcal{C}_B(x^{**})$, then $|\mathcal{C}_A(x^{**})|<|\mathcal{C}_B(x^{**})|$. Note that $\mathcal{C}_A(x)\subseteq\mathcal{C}_B(x)$ holds for any $x\in U$. It follows that $\sum_{x\in U}|\mathcal{C}_A(x)|<\sum_{x\in U}|\mathcal{C}_B(x)|$. This contradicts $\sum_{x\in U}|\mathcal{C}_B(x)|=\sum_{x\in U}|\mathcal{C}_A(x)|$. Hence, for any $x\in U$, $\mathcal{F}_A(x)=\mathcal{F}_B(x)$ and $\mathcal{C}_A(x)=\mathcal{C}_B(x)$ must hold, which means that B is a complementary decision consistent set. □

Different attributes may play different roles in determining the dependency level between condition attributes and label set. In the following discussion, dependency function is used to measure the significance of every condition attribute:

Definition 11. *Let $S=(U,A,L)$ be a multi-label decision table, and let $B\subseteq A$. The inner significance measure of $a\in B$ is defined by*

$$Sig^{inner}(a,B,L)=\gamma_L(B)-\gamma_L(B-\{a\}).$$

Here, $Sig^{inner}(a,B,L)$ reflects the extent to which the dependency level between B and L decreases as a result of removing attribute a from B. Moreover, the indispensability of an attribute can be characterized by the inner significance in the following way:

Proposition 8. *Let $S=(U,A,L)$ be a multi-label decision table, and let $B\subseteq A$ be a complementary decision consistent set. Then, $a\in B$ is an indispensable attribute if and only if $Sig^{inner}(a,B,L)>0$.*

Proof. If $a\in B$ is an indispensable attribute in B, then it follows from Proposition 2 that $B-\{a\}$ is an inconsistent set of S. According to Proposition 7 and Theorem 3, it follows that $\gamma_L(A)=\gamma_L(B)>\gamma_L(B-\{a\})$, i.e., $Sig^{inner}(a,B,L)>0$.

Conversely, if $Sig^{inner}(a,B,L)>0$, then $\gamma_L(B)>\gamma_L(B-\{a\})$. Because B is a consistent set, according to Theorem 3, it follows that $\gamma_L(A)=\gamma_L(B)$. Now, it can be obtained that $\gamma_L(B)>\gamma_L(B-\{a\})$. It follows from Theorem 3 that $B-\{a\}$ is an inconsistent set of S. From Proposition 2, it is clear that a is an indispensable attribute. □

The following property states that the core of A can also be expressed by the inner significance measure:

Proposition 9. *For a multi-label decision table $S = (U, A, L)$, it follows that*

$$CORE(A) = \{a \in A : Sig^{inner}(a, A, L) > 0\}.$$

Proof. By the definition of a complementary decision consistent set, it is known that A is a consistent set. It follows from Proposition 8 that $a \in A$ is an indispensable attribute if and only if $Sig^{inner}(a, A, L) > 0$. According to the definition of the core of A, it follows that $CORE(A) = \{a \in A : Sig^{inner}(a, A, L) > 0\}$. □

Next, the complementary decision reduct will be characterized by the dependency function and the inner significance measure.

Theorem 4. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. If $\gamma_L(A) = \gamma_L(B)$ and $Sig^{inner}(a, B, L) > 0$ for each $a \in B$, then B is a complementary decision reduct of S.*

Proof. The proof is obvious from Theorem 3, Proposition 8, Proposition 2, and Definition 5. □

Next, another attribute significance measure will be presented.

Definition 12. *Let $S = (U, A, L)$ be a multi-label decision table, and let $B \subseteq A$. The outer significance measure of $a \in A - B$ with respect to B is defined by*

$$Sig^{outer}(a, B, L) = \gamma_L(B \cup \{a\}) - \gamma_L(B).$$

Note that $Sig^{outer}(a, B, L)$ is different from $Sig^{inner}(a, B, L)$ because the former is defined for $a \notin B$, whereas the latter is defined for $a \in B$. Furthermore, $Sig^{outer}(a, B, L)$ reflects the extent to which the dependency level between B and L increases as a result of the addition of a to B; this means that the larger the value of $Sig^{outer}(a, B, L)$, the more significant a is. Hence, $Sig^{outer}(a, B, L)$ can be used as a heuristic information to compute a complementary decision reduct.

Based on the above discussion, a heuristic algorithm can be formulated to search for a complementary decision reduct. This procedure is outlined in Algorithm 1.

In Algorithm 1, the first step is to compute $CORE(A)$ according to Proposition 9 in Steps 2-7. Steps 8-19, starting from $E = CORE(A)$, heuristically add to E the attributes that have relatively higher outer significance until a complementary decision consistent set is obtained, i.e., $\gamma_L(E) = \gamma_L(A)$ by Theorem 3. Then, Step 20 sets $RED = E$ and Steps 21-30 remove redundant attributes from RED according to Proposition 8. Finally, by Theorem 4, a reduct is obtained in Step 31.

The next step is to analyze the time complexity of Algorithm 1. The time complexity of Steps 2-7 is $O(|U||A|^2)$, and that of Steps 9-19 is $O(|U||A - E||E|)$. Moreover, the time complexity of Steps 21-30 is $O(|U||RED|^2)$. Hence, the time complexity of Algorithm 1 is $O(|U||A|^2)$.

5. Experiments

This section aims to compare the complementary decision reduct algorithm (CDR) with the two other representative attribute reduction algorithms, namely, the positive region reduct algorithm (PRR) [13] and the Shannon's condition information entropy reduct algorithm (SCER) [43], in terms of the number of selected attributes, the running time to compute one reduct, and six evaluation measures defined below.

Note that all the three algorithms, PRR, SCER, and the proposed algorithm, CDR, use forward greedy search strategies to heuristically compute an attribute reduct. The difference lies in the dependency functions used in the search process. For PRR, the dependency function of condition attribute subset B with respect to decision attribute set D is defined by

$$\gamma_D(B) = \frac{|POS_B(D)|}{|U|}.$$

For SCER, the dependency function is defined by

$$H(D|B) = -\sum_{i=1}^{m} \frac{|X_i|}{|U|} \sum_{j=1}^{n} \frac{|X_i| \cap |Y_j|}{|X_i|} log(\frac{|X_i| \cap |Y_j|}{|X_i|}),$$

where $\{X_1, \cdots, X_m\}$ and $\{Y_1, \cdots, Y_n\}$ are the partitions generated by B and D, respectively.

All experiments were conducted on a server with a 16-core 2.40-GHz Intel Xeon E5-2665 CPU and 32 GB of RAM.

The first step was to collect nine multi-label datasets[2] from different domains, and their properties are listed in Table 3.

[2] The datasets are available at http://mlkd.csd.auth.gr/multilabel.html#Datasets and http://meka.sourceforge.net/#datasets.

Algorithm 1 Heuristic algorithm to search for a single complementary decision reduct

```
Input: a multi-label decision table $S = (U, A, L)$
Output: a complementary decision reduct of $S$
set $E := \emptyset, RED := \emptyset$, $CORE(A) := \emptyset$, and $t := 1$
for $a \in A$ do
    compute $Sig^{inner}(a, A, L) = \gamma_L(A) - \gamma_L(A - \{a\})$
    if $Sig^{inner}(a, A, L) > 0$ then
        set $CORE(A) := CORE(A) \cup \{a\}$
    end if
end for
set $E := CORE(A)$
while $\gamma_L(E) \neq \gamma_L(A)$ do
    choose any $c^* \in A - E$
    compute $Sig^{outer}(c^*, E, L) = \gamma_L(E \cup \{c^*\}) - \gamma_L(E)$
    for $c \in A - E$ do
        compute $Sig^{outer}(c, E, L) = \gamma_L(E \cup \{c\}) - \gamma_L(E)$
        if $Sig^{outer}(c, E, L) > Sig^{outer}(c^*, E, L)$ then
            set $c^* := c$
        end if
    end for
    set $E := E \cup \{c^*\}$
end while
set $RED := E$
while $t$ do
    set $t := 0$
    for $r \in RED$ do
        if $Sig^{inner}(r, RED, L) = 0$ then
            set $t := 1$
            set $RED := RED - \{r\}$
            break
        end if
    end for
end while
return $RED$
```

Table 3: Brief description of selected multi-label datasets

Dataset	Type of attributes	Number of instances	Number of attributes	Number of labels	Domain
Music	numerical	593	72	6	media
Scene	numerical	2407	294	6	media
Yeast	numerical	2417	103	14	biology
Genbase	nominal	662	1185	27	biology
Medical	nominal	978	1449	45	text
LangLog	nominal	1460	1004	75	text
Enron	nominal	1702	1001	53	text
Slashdot	nominal	3782	1079	22	text
Corel5k	nominal	5000	499	374	media

5.1. Evaluation measures

Two groups of measures were employed to evaluate the performance of label set prediction and the performance of label ranking [37]. The first group evaluates the performance of label set prediction and involves two notations:

L_x, the set of true labels of instance x;

$h : U \longrightarrow \mathcal{P}(L)$, the label prediction function, where $h(x)$ is the set of labels predicted by a multi-label classifier h for instance x.

Hamming Loss Hamming Loss is one of the most important multi-label evaluation measures, with a wide range of applications in many studies. It computes the percentage of labels that are predicted incorrectly, i.e., cases in which a label not belonging to the instance is predicted or a label belonging to the instance is not predicted. This measure is defined by

$$HammLoss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|L|}|h(x_i)\Delta L_{x_i}|,$$

where Δ is the symmetric difference between two sets. The smaller the value of $HammLoss(h)$, the better the performance becomes; performance is best when $HammLoss(h) = 0$.

F_1score F_1score is a standard information retrieval measure that combines precision and recall of predictions over N test instances. The measure is defined by

$$F_1score = \frac{1}{N}\sum_{i=1}^{N} F_1score_i = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times |h(x_i) \cap L_{x_i}|}{|h(x_i)| + |L_{x_i}|},$$

where F_1score_i corresponds to the harmonic mean between precision and recall of a prediction for a single instance:

$$Precision_i = \frac{|h(x_i) \cap L_{x_i}|}{|h(x_i)|},$$

$$Recall_i = \frac{|h(x_i) \cap L_{x_i}|}{|L_{x_i}|}.$$

The value of F_1score ranges from 0 to 1; the larger the value, the better the performance becomes.

The second group of measures concerns the performance of label ranking for each instance, based on the real-valued scoring function $f : U \times L \longrightarrow \mathbb{R}$. A successful learning system tends to produce larger scores for labels in L_x than for those not in L_x; in particular, $f(x, l_1) > f(x, l_2)$ for any $l_1 \in L_x$ and $l_2 \notin L_x$. The scoring function f can be transformed into a ranking function $rank_f$ such that if $f(x_i, l_1) > f(x_i, l_2)$, then $rank_f(x_i, l_1) < rank_f(x_i, l_2)$. The following multi-label evaluation measures are used in this paper:

One Error One Error evaluates how many times the top ranked labels are not in the set of relevant labels of instances, as defined by

$$OneError(f) = \frac{1}{N}\sum_{i=1}^{N} [\![\arg\max_{l \in L} f(x_i, l) \notin L_{x_i}]\!],$$

where $[\![\pi]\!]$ equals 1 if π holds and 0 otherwise. The smaller the value of $OneError(f)$, the better the performance becomes.

Coverage Coverage evaluates how far it is necessary, on average, to go down the list of ranked labels to cover all the relevant labels of an instance, as defined by

$$Coverage(f) = \frac{1}{N}\sum_{i=1}^{N} \max_{l \in L_{x_i}} rank_f(x_i, l) - 1,$$

where $rank_f(x_i, l)$ denotes the position of label l in the ordering induced by f. The smaller the value of $Coverage(f)$, the better the performance becomes.

Ranking Loss Ranking Loss evaluates the average fraction of label pairs that are encountered in reverse order for an instance:

$$RankLoss(f) = \frac{1}{N}\sum_{i=1}^{N} \frac{|\{(l,l')|f(x_i,l) \leq f(x_i,l'), (l,l') \in L_{x_i} \times \overline{L_{x_i}}\}|}{|L_{x_i}||\overline{L_{x_i}}|},$$

where $\overline{L_x} = L - L_x$ is the set of irrelevant labels. The smaller the value of $RankLoss(f)$, the better the performance becomes; performance is perfect when $Rankloss(f) = 0$.

Average Precision Average Precision is the average percentage of labels ranked above a particular label l of L_x:

$$AvePrec(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|L_{x_i}|}\sum_{l \in L_{x_i}} \frac{|\{l' \in L_{x_i}|rank_f(x_i,l') \leq rank_f(x_i,l)\}|}{rank_f(x_i,l)}.$$

The larger the value of $AvePrec(f)$, the better the performance becomes; performance is perfect when $AvePrec(f) = 1$.

Note that these six measures evaluate the performance of a multi-learning classifier from different aspects; hence, few algorithms can outperform other algorithms on all these measures.

5.2. Results and Discussion

In the experiments, the numerical attributes, which involve the Music, Scene, and Yeast datasets, were first discretized into several equal-width intervals using a discretization algorithm called FIMUS [34]. Then, the three attribute reduction algorithms were run and their performances were verified based on ML-kNN [50] ($k = 10$) with tenfold cross-validation. Here, the performance in the original space was used as the baseline and is denoted as ORI.

The average numbers of selected attributes and the average running times to compute one reduct for the three algorithms are shown in Tables 4 and 5.

Table 4: Average numbers of selected attributes

Dataset	ORI	PRR	SCER	CDR
Music	72	7.1	7.0	7.3
Scene	294	7.7	7.5	7.8
Yeast	103	9.0	8.1	8.8
Genbase	1185	39.0	28.8	28.5
Medical	1449	59.3	54.5	54.6
LangLog	1004	85.4	26.6	27.2
Enron	1001	96.4	655.5	83.8
Slahdot	1079	306.2	283.5	281.3
Corel5k	499	251.1	210.7	208.7

Table 5: Average running times (in seconds) to compute one reduct

Dataset	PRR	SCER	CDR
Music	20.86	38.54	11.73
Scene	400.95	639.15	221.75
Yeast	150.00	241.54	79.23
Genbase	459.13	393.60	311.80
Medical	5031.80	7690.93	2751.22
LangLog	4218.78	3919.88	1529.96
Enron	11984.09	89113.42	4158.29
Slashdot	245363.68	236033.89	195426.26
Corel5k	442326.15	327186.04	284095.47

From Table 4, it is clear that all the three algorithms could remove some unnecessary attributes. The reducts obtained by CDR were, however, more compact than the others, especially for the Enron dataset.

The running time of CDR was also the shortest among the three algorithms on all datasets, as shown in Table 5. This is not surprising if one inspects the time complexities of the three algorithms, as shown in Table 6. In fact, in PRR and SCER, the computation of the dependency functions must be based on the partitions generated by the indiscernibility relations R_A and R_L, whereas R_A is sufficient for this computation in CDR. Based on the fast partition algorithm proposed in [48], the time complexity of the dependency function in PRR is $O(|U||A| + |U||L|)$ and that in SCER is $O(|U||A| + |U|^2 + |U||L|)$, whereas the time complexity of the dependency function in CDR is only $O(|U||A|)$.

Table 6: Time complexities of the three algorithms

Algorithm	Time complexity
PRR	$O(\|U\|\|A\|^2 + \|U\|\|A\|\|L\|)$
SCER	$O(\|U\|\|A\|^2 + \|U\|^2\|A\| + \|U\|\|A\|\|L\|)$
CDR	$O(\|U\|\|A\|^2)$

Another factor that affects the running time is the number of selected attributes; the smaller the number is, the shorter the running time will be. This can be clearly observed in the Enron dataset. In addition, all the three algorithms were not very efficient for large datasets owing to their high time complexities. For example, the running time of PRR on the Corel5k dataset reached 5 days, and that of CDR reached 3 days.

Tables 7-12 summarize the performances of all the three algorithms on the six evaluation measures defined above. In the experiments, paired t-test was performed using a 0.05 significance level; in Tables 7-12, the symbol "$\oplus$" indicates that CDR is significantly better than the corresponding algorithm on some measure, "$\ominus$" indicates that CDR is significantly worse than the corresponding algorithm, and "$\sim$" indicates that there is no significant difference between CDR and the corresponding algorithm.

Table 7: Comparison of the Hamming Loss measure ($\times 10^1$) (mean±SD)

Dataset	ORI		PRR		SCER		CDR
Music	2.504 ± 0.143	$\ominus$	2.953 ± 0.211	$\sim$	2.997 ± 0.245	$\sim$	2.982 ± 0.188
Scene	1.247 ± 0.072	$\ominus$	1.613 ± 0.053	$\sim$	1.573 ± 0.083	$\sim$	1.631 ± 0.046
Yeast	2.104 ± 0.095	$\ominus$	2.303 ± 0.071	$\sim$	2.301 ± 0.075	$\sim$	2.316 ± 0.082
Genbase	0.046 ± 0.012	$\sim$	0.048 ± 0.028	$\sim$	0.049 ± 0.023	$\sim$	0.052 ± 0.026
Medical	0.153 ± 0.020	$\oplus$	0.149 ± 0.017	$\oplus$	0.141 ± 0.023	$\oplus$	0.135 ± 0.020
LangLog	0.183 ± 0.010	$\sim$	0.183 ± 0.008	$\sim$	0.183 ± 0.009	$\sim$	0.184 ± 0.009
Enron	0.520 ± 0.022	$\oplus$	0.511 ± 0.026	$\oplus$	0.523 ± 0.023	$\oplus$	0.505 ± 0.030
Slashdot	0.521 ± 0.015	$\oplus$	0.455 ± 0.019	$\sim$	0.462 ± 0.012	$\oplus$	0.449 ± 0.014
Corel5k	0.094 ± 0.001	$\oplus$	0.094 ± 0.001	$\sim$	0.094 ± 0.001	$\sim$	0.094 ± 0.001

From Tables 7 to 12, the following observations are clear:

Table 8: Comparison of the F_1score measure (mean±SD)

Dataset	ORI		PRR		SCER		CDR
Music	0.429 ± 0.057	⊖	0.230 ± 0.045	∼	0.235 ± 0.061	∼	0.192 ± 0.071
Scene	0.474 ± 0.039	⊖	0.195 ± 0.014	∼	0.251 ± 0.030	∼	0.192 ± 0.030
Yeast	0.566 ± 0.019	⊖	0.477 ± 0.024	∼	0.475 ± 0.025	∼	0.477 ± 0.019
Genbase	0.954 ± 0.023	∼	0.960 ± 0.027	∼	0.956 ± 0.025	∼	0.946 ± 0.039
Medical	0.589 ± 0.056	⊕	0.616 ± 0.050	⊕	0.656 ± 0.051	⊕	0.665 ± 0.052
LangLog	0.027 ± 0.016	⊖	0.011 ± 0.009	∼	0.011 ± 0.010	∼	0.010 ± 0.012
Enron	0.427 ± 0.030	⊕	0.466 ± 0.025	∼	0.431 ± 0.037	⊕	0.482 ± 0.035
Slashdot	0.057 ± 0.018	⊕	0.255 ± 0.029	∼	0.246 ± 0.029	⊕	0.271 ± 0.025
Corel5k	0.018 ± 0.004	⊖	0.012 ± 0.006	∼	0.008 ± 0.003	∼	0.011 ± 0.005

Table 9: Comparison of the One Error measure (mean±SD)

Dataset	ORI		PRR		SCER		CDR
Music	0.395 ± 0.064	⊖	0.506 ± 0.066	∼	0.528 ± 0.077	∼	0.501 ± 0.075
Scene	0.358 ± 0.033	⊖	0.547 ± 0.033	∼	0.507 ± 0.028	⊖	0.544 ± 0.021
Yeast	0.252 ± 0.028	∼	0.251 ± 0.033	∼	0.252 ± 0.033	∼	0.252 ± 0.034
Genbase	0.014 ± 0.021	∼	0.014 ± 0.020	∼	0.009 ± 0.008	∼	0.011 ± 0.016
Medical	0.249 ± 0.043	⊕	0.237 ± 0.050	∼	0.237 ± 0.057	∼	0.230 ± 0.058
LangLog	0.804 ± 0.043	⊖	0.844 ± 0.031	∼	0.840 ± 0.021	∼	0.858 ± 0.020
Enron	0.304 ± 0.037	⊕	0.284 ± 0.034	∼	0.309 ± 0.041	⊕	0.271 ± 0.027
Slashdot	0.641 ± 0.018	⊕	0.543 ± 0.020	∼	0.552 ± 0.018	∼	0.545 ± 0.023
Corel5k	0.736 ± 0.014	∼	0.719 ± 0.014	∼	0.738 ± 0.018	∼	0.733 ± 0.017

Table 10: Comparison of the Coverage measure (mean±SD)

Dataset	ORI		PRR		SCER		CDR
Music	2.205 ± 0.135	$\ominus$	2.612 ± 0.237	$\sim$	2.570 ± 0.172	$\sim$	2.623 ± 0.290
Scene	0.787 ± 0.058	$\ominus$	1.348 ± 0.110	$\sim$	1.198 ± 0.072	$\ominus$	1.315 ± 0.116
Yeast	6.539 ± 0.173	$\ominus$	6.830 ± 0.172	$\sim$	6.815 ± 0.198	$\sim$	6.811 ± 0.226
Genbase	0.573 ± 0.264	$\sim$	0.516 ± 0.266	$\sim$	0.526 ± 0.258	$\sim$	0.579 ± 0.262
Medical	2.687 ± 0.504	$\sim$	2.776 ± 0.703	$\sim$	2.836 ± 0.741	$\sim$	2.753 ± 0.746
LangLog	15.323 ± 1.552	$\ominus$	16.455 ± 1.110	$\sim$	16.354 ± 1.077	$\sim$	16.368 ± 1.273
Enron	13.070 ± 0.922	$\sim$	13.193 ± 0.973	$\sim$	13.148 ± 0.927	$\sim$	13.132 ± 1.085
Slashdot	4.118 ± 0.207	$\oplus$	3.486 ± 0.204	$\sim$	3.472 ± 0.216	$\sim$	3.451 ± 0.256
Corel5k	114.224 ± 5.584	$\oplus$	112.691 ± 4.813	$\sim$	113.617 ± 5.265	$\sim$	113.244 ± 4.565

Table 11: Comparison of the Ranking Loss measure ($\times 10^1$) (mean±SD)

Dataset	ORI		PRR		SCER		CDR
Music	2.446 ± 0.286	$\ominus$	3.381 ± 0.527	$\sim$	3.245 ± 0.041	$\sim$	3.336 ± 0.602
Scene	1.401 ± 0.108	$\ominus$	2.520 ± 0.207	$\sim$	2.216 ± 0.136	$\ominus$	2.449 ± 0.201
Yeast	1.852 ± 0.137	$\ominus$	2.090 ± 0.139	$\sim$	2.096 ± 0.155	$\sim$	2.089 ± 0.161
Genbase	0.064 ± 0.031	$\sim$	0.053 ± 0.036	$\sim$	0.056 ± 0.037	$\sim$	0.065 ± 0.038
Medical	0.414 ± 0.095	$\sim$	0.427 ± 0.124	$\sim$	0.446 ± 0.143	$\sim$	0.426 ± 0.133
LangLog	1.651 ± 0.225	$\ominus$	1.786 ± 0.171	$\sim$	1.777 ± 0.172	$\sim$	1.775 ± 0.181
Enron	0.921 ± 0.104	$\sim$	0.917 ± 0.112	$\sim$	0.913 ± 0.102	$\sim$	0.900 ± 0.112
Slashdot	1.728 ± 0.101	$\oplus$	1.429 ± 0.089	$\sim$	1.420 ± 0.087	$\sim$	1.417 ± 0.099
Corel5k	1.342 ± 0.064	$\oplus$	1.317 ± 0.056	$\sim$	1.332 ± 0.061	$\sim$	1.326 ± 0.061

Table 12: Comparison of the Average Precision measure (mean±SD)

Dataset	ORI	PRR	SCER	CDR
Music	0.716 ± 0.031 $\ominus$	0.644 ± 0.043 $\sim$	0.637 ± 0.041 $\sim$	0.644 ± 0.049
Scene	0.779 ± 0.019 $\ominus$	0.645 ± 0.024 $\sim$	0.676 ± 0.016 $\ominus$	0.648 ± 0.019
Yeast	0.736 ± 0.020 $\ominus$	0.708 ± 0.019 $\sim$	0.708 ± 0.021 $\sim$	0.707 ± 0.022
Genbase	0.986 ± 0.008 $\sim$	0.987 ± 0.009 $\sim$	0.987 ± 0.005 $\sim$	0.988 ± 0.008
Medical	0.808 ± 0.033 $\sim$	0.814 ± 0.039 $\sim$	0.809 ± 0.042 $\sim$	0.816 ± 0.038
LangLog	0.304 ± 0.036 $\ominus$	0.270 ± 0.022 $\sim$	0.267 ± 0.017 $\ominus$	0.257 ± 0.023
Enron	0.635 ± 0.020 $\oplus$	0.641 ± 0.023 $\oplus$	0.634 ± 0.020 $\oplus$	0.652 ± 0.022
Slashdot	0.500 ± 0.015 $\oplus$	0.574 ± 0.015 $\sim$	0.571 ± 0.013 $\sim$	0.575 ± 0.017
Corel5k	0.246 ± 0.006 $\sim$	0.254 ± 0.007 $\sim$	0.243 ± 0.009 $\sim$	0.247 ± 0.009

- CDR was competitive with PRR on all datasets, whether numerical data or categorical data. For example, CDR was significantly better than PRR according to the Hamming Loss measure on the Medical and Enron datasets. This means that the classifier with the help of CDR was more correct than the others: the original relevant labels were more likely to be predicted as relevant, and the original non-relevant labels were more likely to be predicted as non-relevant. CDR was also significantly better than PRR according to the F_1score measure on the Medical dataset. This means that the prediction of relevant labels by the classifier with the help of CDR was more exact (corresponding to high precision) and more complete (corresponding to high recall) than the other predictions. In addition, CDR was significantly better than PRR according to the Average Precision measure on the Enron dataset; hence, the classifier with the help of CDR could be more effective in predicting the ranking of the relevant labels.

- CDR had better performance than PRR and SCER on the two measures that are relevant to label prediction: the Hamming Loss measure and the F_1score measure. For example, CDR was significantly better than PRR and SCER on the Medical dataset.

- CDR had roughly the same performance as SCER on the four measures that are relevant to label ranking performance for categorical data.

For example, CDR was significantly better than SCER according to the One Error and Average Precision measures on the Enron dataset, whereas the opposite was true with the Average Precision measure on the LangLog dataset.

- CDR was not as good as SCER on the four measures that are relevant to label ranking performance for numerical data. For example, CDR was significantly worse than SCER according to these four measures on the Scene dataset. In fact, all the three algorithms were worse than ORI according to all the measures on numerical data. The reason for this is that all the three algorithms worked with the indiscernibility of instances and, hence, might be inaccurate owing to discretization. Therefore, more reasonable attribute reduction methods specific to numerical data should be considered in the future.

- For the categorical data, CDR was slightly better than ORI according to the two measures that are relevant to label predictive performance, especially the Hamming Loss measure. For example, CDR was significantly better than ORI according to the Hamming Loss measure on four of the datasets. Moreover, CDR had roughly the same performance as ORI on the four measures that are relevant to label ranking performance. For example, CDR was significantly better than ORI according to the four measures on the Slashdot dataset, but not on the LangLog dataset.

In summary, CDR could find more compact reducts than the other methods in the shortest time and, moreover, had better performance than PRR and better label predictive performance than SCER. However, all the three attribute reduction algorithms, including the proposed algorithm, were ineffective in handling numerical data, and because of high time complexity, the three algorithms were also not scalable for large datasets (see Appendix).

6. Conclusions

The complementary decision reduct presented in this paper is a new type of attribute reduct designed for multi-label data, which aims to remove unnecessary attributes while preserving the uncertainties conveyed by labels.

Some of its theoretical properties have been shown here, demonstrating significant advantages of complementary decision reduct in revealing the uncertainties implied in multi-label data. A discernibility matrix-based method and a heuristic algorithm for computing complementary decision reduct have also been proposed. Experiments show that the proposed attribute reduction method not only improves classifier performance for categorical data, but is also competitive with the other two attribute reduction methods on label predictive performance.

In the future, it is interesting to extend the proposed method to deal with numerical data, and to improve the implementation of the proposed method to make it scalable for large datasets [16, 36].

Acknowledgements

The authors would like to thank Professor Witold Pedrycz and the anonymous referees for their valuable comments and suggestions that have significantly improved the quality and presentation of this paper. This work was supported by the National Natural Science Foundation (grant nos. 61272095, 61175067, 71031006, 61303107, and 71201111), the Shanxi Scholarship Council of China (grant no. 2013-014), a Project Supported by National Science and Technology Project of China (no. 2012BAH33B01), and the Natural Science Foundation of the Hebei Education Department (grant no. Z2014106).

Appendix. Experiments on a large dataset

One anonymous reviewer has advised to run the proposed algorithm on the large dataset used in the JRS'12 Data Mining Competition [15], which consists of 10,000 instances with 25,640 attributes and 83 labels. Unfortunately, because of high time complexity, both the proposed algorithm, CDR, and the other two attribute reduction algorithms, PRR and SCER, failed to select valuable attributes on a server with a 16-core 2.80-GHz Intel E5-2680VI CPU and 256 GB of RAM within 10 days.

The recent work by Janusz and Ślęzak [16] showed the significant performance improvements of the attribute reduction algorithms with the help of some randomization techniques and attribute clustering methods. Inspired by that, we speed up the proposed algorithm, CDR, and the other two attribute reduction algorithms, PRR and SCER, by clustering attributes, sampling from clusters, and performing attribute reduction in the sample datasets.

Firstly, the k-means clustering algorithm [14] was employed to automatically cluster the attributes. The cosine distance [26], the Pearson correlation-based distance [30], and the Chebyshev distance [2] were selected as distance measures, respectively. The Euclidean distance was excluded because it is "concentrated" in high-dimensional spaces, i.e., all pairwise distances in high-dimensional spaces were very similar [6]. In addition, the optimal number of clusters was obtained by the cluster validity index [25].

Secondly, conditional attributes were randomly sampled from all the clusters, and the sample sizes were determined by a common approach [19]. The sample datasets corresponding to the three distance measures were denoted as the cosine-sample dataset, the Pearson-sample dataset, and the Chebyshev-sample dataset, respectively.

Finally, the three attribute reduction algorithms were run on the above sample datasets on a server with a 16-core 2.80-GHz Intel E5-2680VI CPU and 256 GB of RAM, and their performances were compared in terms of the number of selected attributes, the running time of computing one reduct, and the six evaluation measures. ML-kNN ($k = 10$) was used to verify the performance of the attribute reduct. The performance in the original dataset was used as the baseline and is denoted as ORI.

The numbers of the selected attributes are shown in Table A.1.

Table A.1: Comparison of the numbers of selected attributes

Sample Dataset	ORI	PRR	SCER	CDR
cosine-sample	25640	422	422	422
Pearson-sample	25640	398	399	399
Chebyshev-sample	25640	396	396	396

From Table A.1, we see that the numbers of selected attributes by the three algorithms were almost same for all sample datasets. Furthermore, we checked the attribute reducts obtained by the three algorithms and found that they were almost same. This may be due to the sparseness of the sample datasets, but it is not assured.

The running times of obtaining one reduct are listed in Table A.2.

It is easy to see from Table A.2 that all the three algorithms were not very efficient for the sample datasets. This may be due to their high time

Table A.2: Comparison of the running times (in hours) to compute one reduct

Sample Dataset	PRR	SCER	CDR
cosine-sample	47.0	46.9	46.1
Pearson-sample	48.2	47.9	47.5
Chebyshev-sample	22.2	22.1	22.0

complexities.

The performances of the compared algorithms on the cosine-sample dataset are shown in Table A.3. For each evaluation measure, the best result is highlighted in boldface. The performances of the compared algorithms on the other two sample datasets are similar and, thus, are omitted.

Table A.3: Comparison of the six evaluation measures on the cosine-sample dataset

Evaluation measure	ORI	PRR	SCER	CDR
Hamming loss	**0.0385**	0.0426	0.0426	0.0426
F_1*score*	**0.3247**	0.1257	0.1257	0.1257
One error	**0.3432**	0.5712	0.5712	0.5712
Coverage	**19.5811**	27.3035	27.3035	27.3035
Ranking loss	**0.0889**	0.1520	0.1520	0.1520
Average precision	**0.5765**	0.4039	0.4039	0.4039

From Table A.3, we see that all the three algorithms were not competitive with ORI on all the measures. This may have been caused by the sampling methods and the attribute clustering methods. Also note that the three algorithms had the same performances on all the measures; this is straightforward, since the three algorithms had the same attribute reduct.

Although a series of experiments were conducted, the effectiveness of the proposed algorithm on large datasets has not been verified owing to high time complexity. Thus, designing efficient attribute reduction algorithms for large and complex datasets is still an important future work.

References

[1] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognition 37 (2004) 1757–1771.

[2] C. Cantrell, Modern Mathematical Methods for Physicists and Engineers, Cambridge University Press, New York, NY, USA, 2000.

[3] D. Chen, S. Zhao, L. Zhang, Y. Yang, X. Zhang, Sample pair selection for attribute reduction with rough set, IEEE Transactions Knowledge and Data Engineering 24 (2012) 2080–2093.

[4] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.

[5] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in Neural Information Processing Systems 14, MIT press, Cambridge, MA, 2002, pp. 681–687.

[6] D. François, V. Wertz, M. Verleysen, The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering 19 (2007) 873–886.

[7] N. Ghamrawi, A. McCallum, Collective multi-label classification, in: O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, W. Teiken (Eds.), Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2005, pp. 195–200.

[8] S. Greco, B. Matarazzo, R. Slowinski, Rough approximation by dominance relations, International Journal of Intelligent Systems 17 (2002) 153–171.

[9] J.W. Grzymała-Busse, Managing Uncertainty in Expert Systems, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[10] J.W. Grzymała-Busse, LERS-A system for learning from examples based on rough sets, in: R. Słowiński (Ed.), Intelligent Decision Support, volume 11 of *Theory and Decision Library*, Springer Netherlands, 1992, pp. 3–18.

[11] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.

[12] Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (2006) 414–423.

[13] X. Hu, N. Cercone, Learning in relational databases: A rough set approach, Computational Intelligence 11 (1995) 323–338.

[14] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (1999) 264–323.

[15] A. Janusz, H.S. Nguyen, D. Ślęzak, S. Stawicki, A. Krasuski, JRS'2012 data mining competition: Topical classification of biomedical research papers, in: J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra, L. Polkowski (Eds.), Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing, RSCTC 2012, volume 7413 of *Lecture Notes in Computer Science*, Springer, Heidelberg, 2012, pp. 422–431.

[16] A. Janusz, D. Ślęzak, Rough set methods for attribute clustering and selection, Applied Artificial Intelligence 28 (2014) 220–242.

[17] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems 141 (2004) 469–485.

[18] R. Jensen, A. Tuson, Q. Shen, Finding rough and fuzzy-rough set reducts with SAT, Information Sciences 255 (2014) 100–120.

[19] J. Jia, X. He, Y. Jin, Principles of Statistics, China Renmin University Publishing, Beijing, China, 2009.

[20] S. Kiritchenko, Hierarchical text categorization and its application to bioinformatics, Ph.D. thesis, University of Ottawa, Ontario, Canada, 2005.

[21] M. Kryszkiewicz, Rough set approach to incomplete information systems, Information Sciences 112 (1998) 39–49.

[22] M. Kryszkiewicz, Comparative study of alternative types of knowledge reduction in inconsistent systems, International Journal of Intelligent Systems 16 (2001) 105–120.

[23] D. Li, B. Zhang, Y. Leung, On knowledge reduction in inconsistent decision information systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12 (2004) 651–672.

[24] Y. Liu, W. Huang, Y. Jiang, Z. Zeng, Quick attribute reduct algorithm for neighborhood rough set model, Information Sciences 271 (2014) 65–81.

[25] B. Mirkin, Reinterpreting the category utility function, Machine Learning 45 (2001) 219–228.

[26] A. Ochiai, Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions, Bulletin of the Japanese Society of Scientific Fisheries 22 (1957) 526–530.

[27] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (1982) 341–356.

[28] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[29] W. Pedrycz, Granular computing: Analysis and Design of Intelligent Systems, Industrial electronics series, Taylor Francis, Boca Raton, 2013.

[30] T. Prasad, S. Ahson, Data mining for bioinformatics — microarray data, Data Mining for Bioinformatics — Microarray Data, Springer Netherlands, Dordrecht, 2009, pp. 77–144.

[31] J. Qian, D. Miao, Z. Zhang, X. Yue, Parallel attribute reduction algorithms using MapReduce, Information Sciences 279 (2014) 671–690.

[32] Y. Qian, J. Liang, W. Pedrycz, C. Dang, Positive approximation: An accelerator for attribute reduction in rough set theory, Artificial Intelligence 174 (2010) 597–618.

[33] B. Qin, ∗-reductions in a knowledge base, Information Sciences 320 (2015) 190–205.

[34] M. Rahman, M. Islam, FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis, Knowledge-Based Systems 56 (2014) 311–327.

[35] J. Read, Scalable multi-label classification, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 2010.

[36] L.S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Ślęzak, J.M. Benítez, Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets", Information Sciences 287 (2014) 68–89.

[37] R.E. Schapire, Y. Singer, BoosTexter: A boosting-based system for text categorization, Machine Learning 39 (2000) 135–168.

[38] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Słowiński (Ed.), Intelligent Decision Support, volume 11 of *Theory and Decision Library*, Springer Netherlands, 1992, pp. 331–362.

[39] D. Ślęzak, Approximate entropy reducts, Fundamenta Informaticae 53 (2002) 365–390.

[40] D. Ślęzak, On generalized decision functions: Reducts, networks and ensembles, in: Y. Yao, Q. Hu, H. Yu, J.W. Grzymala-Busse (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: Proceedings of the 15th International Conference, RSFDGrC 2015, Tianjin, China, volume 9437 of *Lecture Notes in Computer Science*, Springer International Publishing Switzerland, 2015, pp. 13–23.

[41] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multi-label classification of music into emotions, in: J.P. Bello, E. Chew, D. Turnbull (Eds.), Proceedings of the 9th International Conference on Music Information Retrieval, ISMIR 2008, Philadelphia, PA, USA, pp. 325–330.

[42] C. Wang, Q. He, D. Chen, Q. Hu, A novel method for attribute reduction of covering decision systems, Information Sciences 254 (2014) 181–196.

[43] G. Wang, H. Yu, D. Yang, Decision table reduction based on conditional information entropy, Chinese Journal of Computer 25 (2002) 759–766.

[44] X. Wang, E.C.C. Tsang, S. Zhao, D. Chen, D.S. Yeung, Learning fuzzy rules from fuzzy samples based on rough set technique, Information Sciences 177 (2007) 4493–4514.

[45] S. Watanabe, Knowing and Guessing: A Quantitative Study of Inference and Information, Wiley, New York, NY, USA, 1969.

[46] S. Wong, W. Ziarko, On Optimal Decision Rules in Decision Tables, Technical Report, University of Regina. Department of Computer Science, 1985.

[47] W. Wu, M. Zhang, H. Li, J. Mi, Knowledge reduction in random information systems via Dempster-Shafer theory of evidence, Information Sciences 174 (2005) 143–164.

[48] Z. Xu, Z. Liu, B. Yang, W. Song, A quick attribute reduction algorithm with complexity of $\max(O(|C||U|), O(|C|^2|U/C|))$, Chinese Journal of Computers 29 (2006) 391–399.

[49] Y. Yao, Y. Zhao, Attribute reduction in decision theoretic rough set models, Information Sciences 178 (2008) 3356–3373.

[50] M. Zhang, Z. Zhou, Ml-knn: A lazy learning approach to multi-label learning, Pattern Recognition 40 (2007) 2038–2048.

[51] W. Zhang, J. Mi, W. Wu, Knowledge reductions in inconsistent information systems, Chinese Journal of Computers 26 (2003) 12–18.

[52] S. Zhao, X. Wang, D. Chen, E.C. Tsang, Nested structure in parameterized rough reduction, Information Sciences 248 (2013) 130–150.

[53] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.