Inconsistency guided robust attribute reduction

Qu, Yanpeng; Xu, Zheng; Shang, Changjing; Ge, Xiaolong; Deng, Ansheng; Shen, Qiang

Published in:
Information Sciences

DOI:
10.1016/j.ins.2021.08.049

Publication date:
2021

Citation for published version (APA):
Qu, Y., Xu, Z., Shang, C., Ge, X., Deng, A., & Shen, Q. (2021). Inconsistency guided robust attribute reduction. *Information Sciences*, *580*, 69-91. https://doi.org/10.1016/j.ins.2021.08.049

Inconsistency Guided Robust Attribute Reduction

Yanpeng Qu[a,*], Zheng Xu[a], Changjing Shang[b], Xiaolong Ge[a], Ansheng Deng[a], Qiang Shen[b]

[a] *Information Science and Technology College, Dalian Maritime University, Dalian, 116026, China*
[b] *Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK*

Abstract

Attribute reduction (AR) plays an important role in reducing irrelevant and redundant domain attributes, while maintaining the underlying semantics of retained ones. Based on Earth Mover's Distance (EMD), this paper presents a robust AR algorithm from the perspective of minimising the inconsistency between the discernibility of the reduct and the entire original attribute set. Due to the susceptibility of the inconsistency gauger to noisy information, a strategy for instance denoising is also proposed by detecting abnormal local class distributions with regard to the global class distribution. With such a pretreatment process for AR, the robustness of the reduct found is significantly improved, as testified by systematic experimental investigations. The experimental results demonstrate that the reduct gained by the proposed approach generally outperforms those attained by the application of popular, state-of-the-art AR techniques, in terms of both the size of attribute reduction and the classification results using the reduced attributes.

Keywords: Attribute reduction, Inconsistency, Earth Mover's Distance, Classification, Robustness.

1. Introduction

Attribute reduction (AR) is currently one of the most significant approaches to data preprocessing. By removing the redundant and irrelevant attributes in a given problem, AR helps preserve the information of data

*Corresponding author.
Email address: yanpengqu@dlmu.edu.cn (Yanpeng Qu)

while improving prediction performance in terms of speed and data understanding. It may even help increase accuracy in problem-solving through removing noisy attributes. In general, AR techniques can be divided into three categories [30]: wrapper, embedded and filter. A wrapper method [17] usually selects attributes by evaluating the associated predictive ability. An embedded method is to interlace machine learning algorithms [18] and AR into one indivisible framework. The key factor of the filter approach is to design an evaluation function according to certain criteria to evaluate each attribute and select the attributes that meet the conditions. This evaluation function can be designed from various perspectives, such as information theory [28], loss function with regularisation [16], discernibility [6] and consistency [7].

A concept that is often addressed in AR techniques, especially for the consistency-based methods, is data or dataset consistency. A consistent dataset is one that does not entail a contradiction between its condition attributes and decision attribute. Consistency in a dataset may be measured following the principle that no two instances may have the same value on all predicting attributes if they are associated with a different concept or class [7]. In practice, consistency is usually evaluated via the differences between the data partitions over the condition attributes in relation to the decision attribute. A smaller difference between the distributions of any two partitions will indicate a higher consistency between the discrimination abilities of the two sets of the conditional attributes regarding the common decision attribute. The effect of utilising a consistency measure in implementing AR methods is examined in [9], in conjunction with the application of different search strategies. Consistency is imposed by finding expressions in the diversity of the classes in the conditional attribute set where all samples share identical attribute values. Also, in [37], consistency-based attribute selection is developed by the use of greedy least squares regression. Furthermore, the concept of consistency and its measure have been utilised to determine the degree of how the decision attribute depends on the set of condition attributes, within typical algorithms that are based on rough sets [19],

Instead of exploiting consistency, the notion of inconsistency has also been employed to detect and remove noisy instances or redundant and irrelevant attributes from datasets. For example, variable precision rough sets (VPRS) [32] have been used to reinforce the classic rough set-based AR techniques via measuring the degree of inconsistency in the dataset. Nevertheless, VPRS only reduces the sensitivity of the model to noise according

to a given confidence threshold, rather than in response to directly measured noisy information [12]. In [7], a heuristic is used for selecting neighbouring inconsistent data pairs and through which, an efficient method for finding a attribute reduct is introduced. Unfortunately, the existence of noisy samples may also lead to inaccurate inconsistent sample pairs.

Another technique in the literature is to exploit the inconsistency degree as the fitness function, thereby enabling a genetic algorithm to perform AR [8]. Again, since the state of an individual may be disturbed by noise, such a genetic algorithm may suffer from unstable final optimal solutions. An efficient AR algorithm is proposed in [5] to evaluate the inconsistency of feature subsets at group level. However, the stability of the groups of attributes may also be reduced by noise. By measuring the inconsistency between the partitions of the dataset induced by attribute subsets, the earth mover's distance (EMD) [34, 35] (also known as discrete Wasserstein distance [38]) has been adopted for implement AR. Yet, similar to the other techniques mentioned above, the partition of the dataset induced by the reduct may be impeded by the existence of noisy information. In short, whilst a measure of inconsistency may offer a complementary method to the consistency-based approach, capable of detecting redundant information, given the nature of inconsistency, existing algorithms are susceptible to noise. To combat the potential adverse impact caused by noise, a robust AR method supported with a preprocessing mechanism of data denoising is presented in this paper, improving the quality of AR.

Particularly, a metric to measure inconsistency based on EMD is exploited in this work, in support of the evaluation of an attribute reduct. By using this EMD-based inconsistency metric, an attribute reduct is expected to be produced that will have an identical discernibility to the one attainable by the entire set of the original condition attributes, with regard to the decision attribute. In developing this work, the mechanism to detect any noisy data instance is based on the (rational) assumption that a local class distribution within the neighbourhood [21, 22] of an abnormal sample is allowed to be inconsistent with the global class distribution in the universe of discourse. As a result, two indicators for local and global class distributions are proposed to help determine such noise. With any contradiction amongst the samples being resolved, the quality of the corresponding reduct generated by the subsequent AR procedure becomes more robust and reliable than otherwise.

The resulting robust AR approach is fully implemented, supported with systematic experimental validation and evaluation. To facilitate comparative

analyses, experimental studies are conducted in reference to state-of-the-art and powerful attribute selection methods, including: ASNAR [13], VPRS [32], FKD [35], GBNRS [33], RelieF [31], and SPDTRS [29]. Furthermore, the resultant reducts are applied by the following four different classifiers: Logistic [36], SVM [16], Adaboost with J48 (AdaJ48) [14], and random forest (RF) [20], respectively. The comparative results demonstrate that the proposed approach outperforms the rest, returning reducts that ensure a high classification accuracy across a range of benchmark datasets.

The contribution, innovation and highlights of this work are outlined as follows.

- *Contribution*: Based on EMD, an AR algorithm is developed, organically integrating a data inconsistency measure and a data denoising strategy to guarantee the robustness of the resulting attribute reduct. The strengths of this algorithm are verified from both theoretical and experimental perspectives.

- *Innovation*: Two novel computational mechanisms are offered: 1) to evaluate the significance of a set of attributes via the distribution of the emerging partitions of the universe, and 2) to detect noisy data via the abnormal distribution in its neighbourhood.

- *Highlight*: During the process of refining a dataset, the inconsistency measure in action behaves in the same way as information entropy from two perspectives: reaching its maximum when the partition induced by an attribute subset follows a uniform distribution, and showing monotonicity with respect to the number of equally likely events. The significant effects of the proposed denoising strategy in relieving the impact of noisy data on AR are empirically demonstrated.

The remainder of this paper is structured as follows. In Section 2, the concept of EMD is reviewed and a new inconsistency metric based on EMD is presented. Using the proposed inconsistency metric, Section 3 presents the strategies and their implementations for denoising data and reducing conditional attributes. In Section 4, experimental results are analysed in comparison with popular, state-of-the-art attribute selection techniques. Section 5 concludes the paper with a brief discussion about further research.

2. Earth Movers' Distance and consistency

In this section, a review of EMD and an EMD-based measure to gauge the inconsistency amongst the samples in a dataset are presented. Particularly, the concept of consistency discussed herein reflects the agreement of the decisions with respect to a set of conditional attributes, within certain given data instances.

2.1. Earth Mover's Distance

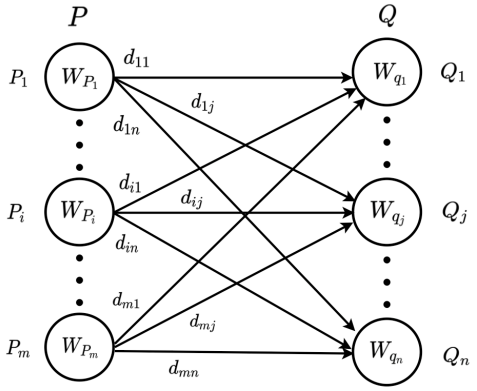


Figure 1: Diagram of Earth mover's distance

In statistics, Earth Mover's Distance (EMD) is a measure of the distance between two discrete probability distributions. The calculation of EMD can be treated as a transportation problem in linear programming [34, 35]. Let $P = \{(p_1, w_{p1}), (p_2, w_{p2}) \dots (p_m, w_{pm})\}$ and $Q = \{(q_1, w_{q1}), (q_2, w_{q2}) \dots (q_n, w_{qn})\}$ represent two distributions (or signatures in pattern recognition) in Fig. 1, where p_i $(i = 1, \dots, m)$ is the i-th cluster of P; q_j $(j = 1, \dots, n)$ is the j-th cluster of Q; and w_{pi} and w_{qj} denote the weight of cluster p_i and that of q_j respectively, and d_{ij} is the distance (or the dissimilarity degree) between p_i and q_j.

As a linear programming problem, EMD measures the distance between P and Q by finding a flow $[f_{ij}]$ in an effort to minimise the total cost of transforming distribution P to distribution Q. Such a problem can be formulated as the following optimisation problem:

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \tag{1}$$

s.t.

$$f_{ij} \geq 0,\ i = 1 \ldots m,\ j = 1 \ldots\ n, \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{pi},\ i = 1 \ldots m, \tag{3}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{qi},\ j = 1 \ldots\ n, \tag{4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\{\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}\}. \tag{5}$$

Constraint (2) guarantees that the flow from p_i to q_j cannot be negative, because a negative quantity cannot be moved from P to Q. Constraints (3) and (4) impose that p_i cannot provide more than its weight and q_i cannot accept more than its weight. Constraint (5) means that if the sum of w_{pi} in distribution P across all i and that of w_{qi} in distribution Q across all j are not equal, the total amount of movement cannot exceed the amount that can be provided or accepted. Resolving the above optimisation problem as stated in expression (1) provides the required EMD measure, denoted by d_E hereafter, which is defined as the work normalised by the total flow:

$$d_E(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}. \tag{6}$$

It has been shown [25] that EMD is a metric that can meet the following three conditions:

- Nonnegativity: $d_E(P, Q) \geq 0$;
- Symmetry: $d_E(P, Q) = d_E(Q, P)$;
- Triangle inequality: $d_E(P, Q) + d_E(Q, R) \geq d_E(P, R)$, when $\sum_{i=1}^{n} w_{pi} = \sum_{j=1}^{m} w_{qj}$.

2.2. Measuring inconsistency with EMD

Let $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$ be an information system, where $\mathbb{U}$ is a finite nonempty set of data instances or objects; $\mathbb{C}$ is a finite nonempty set of nominal condition attributes; D is a decision attribute; V is the value domain

of $a \in \mathbb{C} \cup \{D\}$; and $\phi : \mathbb{U} \times \mathbb{C} \cup \{D\} \to V$ is a value mapping. Given a sample x and an attribute subset $A \subseteq \mathbb{C}$, the agreement of x to each sample in $\mathbb{U}$ with respect to A can be represented as a matrix $[s_{ij}^A(x)]$ whose cell can be generically denoted by

$$s_{ij}^A(x) = \begin{cases} 1, & \phi(x, a_j) = \phi(x_i, a_j) \\ 0, & otherwise \end{cases}, \ i = 1, \ldots, |\mathbb{U}|, \ j = 1, \ldots, |A|, \tag{7}$$

where x_i is the i-th sample in $\mathbb{U}$ and a_j is the j-th attribute in A.

Summing up each column of $[s_{ij}^A(x)]$, namely

$$sum_i^A(x) = \sum_{j=1}^{|A|} s_{ij}^A(x), \ i = 1, \ldots, |\mathbb{U}|, \tag{8}$$

a relation on A can be introduced as follows:

$$E_A = \{(x, x_i) \in \mathbb{U}^2 | sum_i^A(x) = |A|\}. \tag{9}$$

Note that if $sum_i^A(x) = |A|$ it means that x and x_i are identical regarding the attribute subset A. Thus, E_A is an equivalence relation which meets the following three properties:

- Reflexivity: $(x, x) \in E_A$;
- Symmetry: if $(x, x_i) \in E_A$ then $(x_i, x) \in E_A$;
- Transitivity: if $\exists x_i$ and x_k, s.t., $(x, x_i) \in E_A$ and $(x_i, x_k) \in E_A$, then $(x, x_k) \in E_A$.

Table 1: Exemplar dataset

Objects	a	b	c	D
x_1	0	2	2	1
x_2	1	0	0	2
x_3	2	1	0	0
x_4	0	1	2	2
x_5	2	1	0	0
x_6	2	0	0	2

Importantly, with the use of this equivalence relation E_A, $\mathbb{U}$ may be partitioned into equivalence classes, denoted by $\mathbb{U}/E_A$. In particular, for the empty set $\emptyset$, $\mathbb{U}/\emptyset$ is set to $\mathbb{U}$.

Example 1. *To illustrate the concepts involved, a simple dataset is given in Table 1, consisting of three conditional attributes a, b, c, one decision attribute D, and six objects.*

Suppose that $A = \{a, c\}$*, according to Eqs.* (7) *and* (8)*, the agreement matrices per instance can be obtained as follows:*

$$[s_{ij}^A(x_1)] = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^T \Rightarrow (x_1, x_1), (x_1, x_4) \in E_A,$$

$$[s_{ij}^A(x_2)] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}^T \Rightarrow (x_2, x_2) \in E_A,$$

$$[s_{ij}^A(x_3)] = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}^T \Rightarrow (x_3, x_3), (x_3, x_5), (x_3, x_6) \in E_A,$$

$$[s_{ij}^A(x_4)] = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^T \Rightarrow (x_4, x_1), (x_4, x_4) \in E_A,$$

$$[s_{ij}^A(x_5)] = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}^T \Rightarrow (x_5, x_3), (x_5, x_5), (x_5, x_6) \in E_A,$$

$$[s_{ij}^A(x_6)] = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}^T \Rightarrow (x_6, x_3), (x_6, x_5), (x_6, x_6) \in E_A.$$

Due to the properties of symmetry and transitivity of E_A*,* $\{x_1, x_4\}$*,* $\{x_2\}$ *and* $\{x_3, x_5, x_6\}$ *are three equivalence classes in* $\mathbb{U}$ *regarding the attribute subset A. Thus,* $\mathbb{U}/E_A = \{\{x_1, x_4\}, \{x_2\}, \{x_3, x_5, x_6\}\}$.

Without losing generality, suppose that given two attribute subsets $A, B \subseteq \mathbb{C} \cup \{D\}$, the respective partitions generated by A and B are

$$\mathbb{U}/E_A = \{X_{A_1}, X_{A_2}, \ldots, X_{A_i}, \ldots, X_{A_m}\}$$

and

$$\mathbb{U}/E_B = \{X_{B_1}, X_{B_2} \ldots, X_{B_j}, \ldots, X_{B_n}\}.$$

Following the knowledge distance measure as introduced in [34], the degree of inconsistency measured on the basis of EMD between $\mathbb{U}/E_A$ and $\mathbb{U}/E_B$

can be defined by

$$d_I(\mathbb{U}/E_A, \mathbb{U}/E_B) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}}{|\mathbb{U}|}, \tag{10}$$

where

$$f_{ij} = |X_{A_i} \cap X_{B_j}|,\ i = 1 \ldots m,\ j = 1 \ldots\ n, \tag{11}$$

$$d_{ij} = 1 - \frac{|X_{A_i} \cap X_{B_j}|}{|X_{A_i} \cup X_{B_j}|},\ i = 1 \ldots m,\ j = 1 \ldots\ n. \tag{12}$$

with f_{ij} representing the number of samples moved from X_{A_i} to X_{B_j}, and d_{ij} being the Jaccard distance between X_{A_i} and X_{B_j}. Similar to EMD, d_I also fulfills nonnegativity, symmetry and trigonometric inequality as a metric on $\mathbb{U}$. It indicates the cost of transforming one discrete distribution into another discrete distribution. A smaller value of d_I implies a slighter difference between the two corresponding discrete distributions. That is, these two distributions share more information with each other. This can be illustrated with the following example.

Example 2. *Given Table 1, let* $B = \{b\}$, $C = \{c\}$ *and* $B' = B \cup \{D\}$, $C' = C \cup \{D\}$. *The respective partitions of* $\mathbb{U}$ *induced by* B, B', C *and* C' *are*

$$\mathbb{U}/E_B = \{\{x_1\}, \{x_2, x_6\}, \{x_3, x_4, x_5\}\},$$

$$\mathbb{U}/E_{B'} = \{\{x_1\}, \{x_2, x_6\}, \{x_3, x_5\}, \{x_4\}\},$$

$$\mathbb{U}/E_C = \{\{x_1, x_4\}, \{x_2, x_3, x_5, x_6\}\},$$

and

$$\mathbb{U}/E_{C'} = \{\{x_1\}, \{x_2, x_6\}, \{x_3, x_5\}, \{x_4\}\}.$$

From this it can be derived that

$$d_I(\mathbb{U}/E_B, \mathbb{U}/E_{B'}) = \frac{1}{6} \times (2 \times \frac{1}{3} + 1 \times \frac{2}{3}) = \frac{2}{9},$$

and

$$d_I(\mathbb{U}/E_C, \mathbb{U}/E_{C'}) = \frac{1}{6} \times (1 \times \frac{1}{2} + 1 \times \frac{1}{2} + 2 \times \frac{2}{4} + 2 \times \frac{2}{4}) = \frac{1}{2}.$$

The transforming process that this example involves is shown in Fig. 2. It can be seen that from $\mathbb{U}/E_B$ to $\mathbb{U}/E_{B'}$, $\{x_3, x_4, x_5\}$ is divided into $\{x_3, x_5\}$ and $\{x_4\}$, and that from $\mathbb{U}/E_C$ to $\mathbb{U}/E_{C'}$, $\{x_1, x_4\}$ and $\{x_2, x_3, x_5, x_6\}$ are divided into $\{x_1\}, \{x_4\}$, and $\{x_2, x_6\}, \{x_3, x_5\}$, respectively. This indicates that whilst $\mathbb{U}/E_{B'}$ and $\mathbb{U}/E_{C'}$ are both refinements of $\mathbb{U}/E_B$ and $\mathbb{U}/E_C$ induced by D, the cost of transforming $\mathbb{U}/E_B$ to $\mathbb{U}/E_{B'}$ is less than that from $\mathbb{U}/E_C$ to $\mathbb{U}/E_{C'}$. That is, the difference (d_I) between $\mathbb{U}/E_B$ and $\mathbb{U}/E_{B'}$ is less than that between $\mathbb{U}/E_C$ and $\mathbb{U}/E_{C'}$. Such an observation demonstrates the rationality for introducing d_I as defined above.

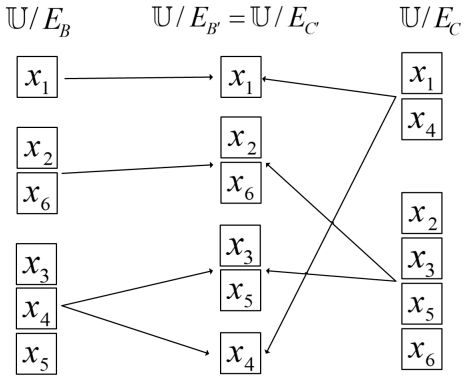


Figure 2: Example graph of inconsistency degree

Regarding a universe $\mathbb{U}$ of the nature as described above, the following two theorems hold for d_I.

Theorem 1. *Given an information system* $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$ *and* $\forall B \subseteq \mathbb{C}$, $\max d_I(\mathbb{U}, \mathbb{U}/E_B) = 1 - \frac{1}{|\mathbb{U}/E_B|}$.

Proof: Suppose that $\mathbb{U}/E_B = \{X_{B_1}, \ldots, X_{B_p}\}$. According to the definition of d_I,

$$d_I(\mathbb{U}, \mathbb{U}/E_B) = \frac{1}{|\mathbb{U}|} \sum_{i=1}^{p} |X_{B_i}|(1 - \frac{|X_{B_i}|}{|\mathbb{U}|}). \tag{13}$$

Then, $\max d_I(\mathbb{U}, \mathbb{U}/E_B)$ is equal to the quadric programming problem

$$\min \frac{1}{|\mathbb{U}|} \sum_{i=1}^{p} |X_{B_i}|(\frac{|X_{B_i}|}{|\mathbb{U}|} - 1), \tag{14}$$

s.t.

$$\sum_{i=1}^{p} |X_{B_i}| - |\mathbb{U}| = 0. \tag{15}$$

The resulting Lagrangian function is

$$L(|X_{B_1}|, \ldots, |X_{B_p}|, \lambda) = \sum_{i=1}^{p} |X_{B_i}|(\frac{|X_{B_i}|}{|\mathbb{U}|} - 1) + \lambda(\sum_{i=1}^{p} |X_{B_i}| - |\mathbb{U}|). \tag{16}$$

By solving

$$\begin{cases} \nabla_{|X_{B_i}|} L(|X_{B_1}|, \ldots, |X_{B_p}|, \lambda) = \frac{2|X_{B_i}|}{|\mathbb{U}|} - 1 + \lambda = 0, i = 1, \ldots, p, \\ \sum_{i=1}^{p} |X_{B_i}| - |\mathbb{U}| = 0, \end{cases} \tag{17}$$

there are $|X_{B_1}| = \cdots = |X_{B_p}| = \frac{|\mathbb{U}|}{p}$ and $\max d_I(\mathbb{U}, \mathbb{U}/E_B) = 1 - \frac{1}{p} = 1 - \frac{1}{|\mathbb{U}/E_B|}$. □

This property demonstrates that d_i is related to the classical information entropy measure, it reaches its maximum when the partition induced by an attribute subset follows an even distribution. For instance, if $|\mathbb{U}/E_B| = 2$, the graph of $d_I(\mathbb{U}, \mathbb{U}/E_B)$ can be shown as per Fig. 3.

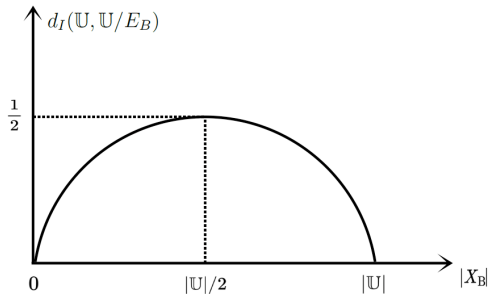


Figure 3: Exemplar graph of inconsistency degree

Theorem 2. *Given an information system* $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$, $\mathbb{U}/E_B = \{X_{B_1}, \ldots, X_{B_p}\}$ *and* $\mathbb{U}/E_{B'} = \{X_{B'_1}, \ldots, X_{B'_q}\}$ *are two partitions of* $\mathbb{U}$ *induced by* B *and* $B' \subseteq \mathbb{C}$, *respectively. If* $\forall X_{B_j} \in \mathbb{U}/E_B$, $\exists \Theta_j \subseteq \{1, \ldots, q\}$, *such that* $|X_{B_j}| = \sum_{i \in \Theta_j} |X_{B'_i}|$, *where* $\cup_{j=1}^{p} \Theta_j = \{1, \ldots, q\}$ *and* $\Theta_j \cap \Theta_l = \emptyset$, *for* $j \neq l$, *then* $d_I(\mathbb{U}, \mathbb{U}/E_B) < d_I(\mathbb{U}, \mathbb{U}/E_{B'})$.

Proof: For simplicity, assume that $q = p{+}1$, $\Theta_1 = \{1, 2\}$ and $\Theta_j = \{j{+}1\}$, i.e., $|X_{B_1}| = |X_{B'_1}| + |X_{B'_2}|$ and $|X_{B_2}| = |X_{B'_3}|, \ldots, |X_{B_p}| = |X_{B'_q}|$. Note that other more complicated cases can be translated into this one.

According to the definition of d_I, it follows that

$d_I(\mathbb{U}, \mathbb{U}/E_B) - d_I(\mathbb{U}, \mathbb{U}/E_{B'})$

$= \frac{1}{|\mathbb{U}|}(|X_{B_1}| \cdot \frac{|\mathbb{U}|-|X_{B_1}|}{|\mathbb{U}|} - |X_{B'_1}| \cdot \frac{|\mathbb{U}|-|X_{B'_1}|}{|\mathbb{U}|} - |X_{B'_2}| \cdot \frac{|\mathbb{U}|-|X_{B'_2}|}{|\mathbb{U}|})$

$= \frac{1}{|\mathbb{U}|}\big((|X_{B'_1}| + |X_{B'_2}|) \cdot \frac{|\mathbb{U}|-(|X_{B'_1}|+|X_{B'_2}|)}{|\mathbb{U}|} - |X_{B'_1}| \cdot \frac{|\mathbb{U}|-|X_{B'_1}|}{|\mathbb{U}|} - |X_{B'_2}| \cdot \frac{|\mathbb{U}|-|X_{B'_2}|}{|\mathbb{U}}|\big)$

$= \frac{1}{|\mathbb{U}|}\big((|X_{B'_1}|(1 - \frac{|X_{B'_1}|+|X_{B'_2}|}{|\mathbb{U}|} - 1 + \frac{|X_{B'_1}|}{|\mathbb{U}|}) + |X_{B'_2}|(1 - \frac{|X_{B'_1}|+|X_{B'_2}|}{|\mathbb{U}|} - 1 + \frac{|X_{B'_2}|}{|\mathbb{U}|})\big)$

$= \frac{1}{|\mathbb{U}|}(-\frac{|X_{B'_1}||X_{B'_2}|}{|\mathbb{U}|} - \frac{|X_{B'_2}||X_{B'_1}|}{|\mathbb{U}|}) < 0$ □

This theorem demonstrates that during the process of refining $\mathbb{U}$, the inconsistency level between $\mathbb{U}$ and its refinement is growing monotonically. This monotonicity of d_I is the same as that of information entropy with respect to the number of equally likely events [27]. Thus, the proposed metric d_I is capable of revealing the inconsistency or uncertainty hidden in an information system.

3. Robust Attribute Reduction

As d_I is able to gauge the transformation degree of a partition of the universe to its refinement, this metric can be employed to explore the possibility of using an attribute subset to represent the original entire attribute set. However, the inconsistency measures amongst the instances may be impeded by the existence of noisy information. In order to perform the search for an effective reduct with noise in the instances diminished, a robust AR algorithm is proposed in this section.

3.1. Instance denoising

Let $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$ be an information system. Given a sample $x \in \mathbb{U}$, an attribute subset $A \subseteq \mathbb{C}$ and a parameter k, suppose that $\Theta_k(x)$ is the set of the indices used to represent the k-nearest neighbours of x. The neighbouring samples concerned may be chosen by the use of a certain metric, such as Hamming distance. In general, within a set of neighbouring samples it is possible that more than one sample is of full similarity or identical to x on their conditional parts. For such neighbouring instances, those in the same class as the given sample x are of particular interest and are therefore, to

be found with priority in the process of searching for the neighbourhood of x. To reflect this intuition, the agreement of x to the instances in the same class as x can be identified by

$$sc_{local}(x) = \frac{\sum_{i \in \Theta_k(x)} sum_i^A(x)}{k * |A|}, \ \phi(x, D) = \phi(x_i, D). \tag{18}$$

Here, $sum_i^A(x)$ is the sum of the i-th column of $[s_{ij}^A(x)]$; $\sum_{i \in \Theta_k(x)} sum_i^A(x)$ ($\phi(x, D) = \phi(x_i, D)$) denotes the total amount of the attributes, on which x and its nearest neighbours in the class as x are identical; $|A|$ is the cardinality of the attribute set A; and k is the number of nearest neighbours of x.

Accordingly, the agreement of x to the instances belonging to any of the classes that are different from the class of x can be defined by

$$dc_{local}(x) = \frac{\sum_{i \in \Theta_k(x)} sum_i^A(x)}{k * |A|}, \ \phi(x, D) \neq \phi(x_i, D). \tag{19}$$

As such, Eqs. (18) and (19) reflect the local consistency and inconsistency degree between x and its k-nearest neighbours, respectively. Based on these two measures, the outlier degree η_{local} of x within its neighbourhood can be introduced and evaluated by

$$\eta_{local}(x) = dc_{local}(x) - sc_{local}(x). \tag{20}$$

The process of calculating the local outlier degree is illustrated in Fig. 4. In this neighbourhood of x where $k = 6$, the agreement of x to those instances within the same class (as x), i.e., class D_1, is calculated with the samples in the dashed box, and the agreement of x to the instances belonging to classes that are different from the class of x, here, the single class D_2, is calculated with the samples in the black box. As shown in Fig. 4, intuitively, if x has more neighbours from its own class, it is less likely to be a noise sample.

In contrast to this local outlier degree of x, global distinction between a class and the others can also be introduced as follows. Given an information system $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$, suppose that $\mathbb{U}$ is composed of L classes: $D_1, \ldots, D_L$, and that the set of the indices of the instances of D_l is denoted by Θ_l. The agreement of D_p to D_q with respect to an attribute subset $A \subseteq \mathbb{C}$

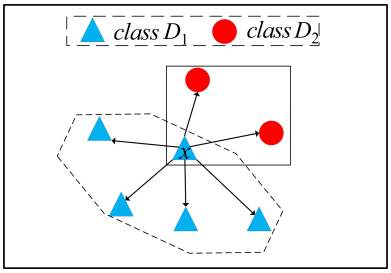


Figure 4: Local outlier degree

can be defined by

$$S(D_p \to D_q) = \frac{\sum_{j\in\Theta_p}\sum_{i\in\Theta_q} sum_i^A(x_j)}{|D_p| * |D_q|}. \tag{21}$$

This makes intuitive sense because $\frac{\sum_{i\in\Theta_q} sum_i^A(x_j)}{|D_q|}$ reflects the average agreement of a sample in D_q with respect to x_j. Then $S(D_p \to D_q)$ represents the average agreement of a sample in D_q with respect to all samples in D_p. In particular, $S(D_q \to D_q)$ indicates the consistency of D_q to itself.

Note that since

$$\sum_{j\in\Theta_p}\sum_{i\in\Theta_q} sum_i^A(x_j) = \sum_{i\in\Theta_q}\sum_{j\in\Theta_p} sum_i^A(x_j),$$

it follows that

$$S(D_p \to D_q) = S(D_q \to D_p).$$

Using the above agreement measure, the global agreement degree of those classes that are distinguished from D_p with respect to an attribute subset $A \subseteq \mathbb{C}$ can be determined by

$$dc_{global}(D_p) = \frac{\sum_{q\neq p}^{L} S(D_q \to D_q)}{(L-1) * |A|},\ p = 1, \ldots L. \tag{22}$$

That is, it measures the distinction degree from class D_q to class D_p globally. Similarly, the agreement of other classes to D_p can be defined by

$$sc_{global}(D_p) = \frac{\sum_{q\neq p}^{L} S(D_q \to D_p)}{(L-1) * |A|},\ p = 1, \ldots L. \tag{23}$$

With the use of Eqs. (21), (22) and (23), the notion of divergence degree η_{global} of a class D_p to which x belongs can be defined by

$$\eta_{global}(D_p) = dc_{global}(D_p) - sc_{global}(D_p). \tag{24}$$

Following Fig. 4, the process of calculating the global divergence degree is illustrated in Fig. 5. In particular, Steps (a), (b) and (c) jointly depict the intermediate example steps of calculating the global agreement with respect to the class D_2 which is distinguished from the class D_1; while Steps (d), (e) and (f) illustrate the course of calculating the class agreement of D_2 to D_1.

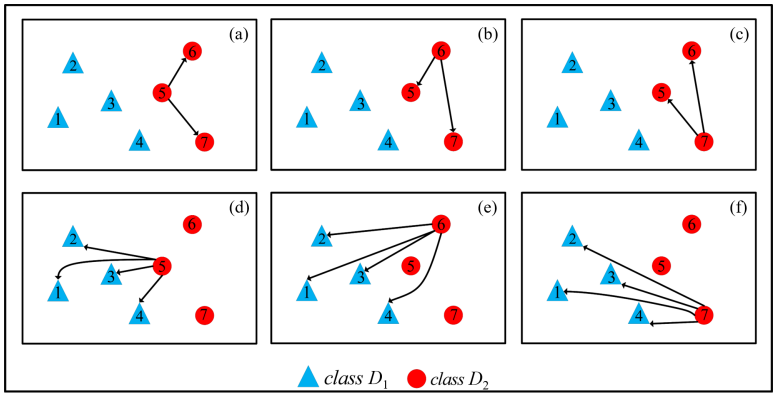


Figure 5: Global divergence degree

Example 3. *To explain the calculation process illustrated in Fig. 5 in more detail, assume that* $sum_i^A(x_j)$, $i = 5, 6, 7$, $j = 1, \ldots, 7$ *and* $|A| = 2$, *are given as listed in Table 2.*

Table 2: $sum_i^A(x_j)$, $i = 5, 6, 7$, $j = 1, \ldots, 7$ and $|A| = 2$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_5	0	1	1	1	2	2	1
x_6	0	1	1	1	2	2	1
x_7	0	0	0	1	1	1	2

According to Eqs. (21), (22), *and* (23),

$$dc_{global}\ (D_1) = \frac{\frac{(2+2+1)+(2+2+1)+(1+1+2)}{3\times 3}}{(2-1)\times 2} = \frac{7}{9},$$

$$sc_{global}\ (D_1) = \frac{\frac{(0+1+1+1)+(0+1+1+1)+(0+0+0+1)}{4\times 3}}{(2-1)\times 2} = \frac{7}{24}.$$

Then,

$$\eta_{global}\ (D_1) = dc_{global}\ (D_1) - sc_{global}\ (D_1) = \frac{105}{216}.$$

Using the computed local and global divergence degrees, an instance $x \in D_p$ will be deemed as a noise data and should therefore be removed from $\mathbb{U}$, if

$$\eta_{local}(x) > \eta_{global}(D_p). \tag{25}$$

Summarising the above, the method of removing noise data is constructed as shown in Alg. 1.

Algorithm 1 Data denoising

Input:

$I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$, an information system;

k, number of nearest neigbours.

Output: $\mathbb{U}'$, denoised dataset.

```
1: U' ← ∅;
2: for ∀D_p ⊂ U, p = 1, ... L
3:     for ∀x ∈ D_p
4:         if η_local(x) > η_global(D_p)
5:         U' ← (U − x).
6:         end
7:     end
8: end
9: Return U'
```

3.2. Attribute reduction

Following Alg. 1, those instances which suffer from abnormal distribution in its neighbourhood are removed from $\mathbb{U}$. On such a denoised dataset, an AR method based on d_I can then be constructed by performing two computational procedures as detailed below.

The first is one for reduct search. Given an information system $I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$, let A be a subset of $\mathbb{C}$. As the size of A increases, each item within the partition $\mathbb{U}/E_A$ tends to become finer. Thus, the inconsistency between the distributions of $\mathbb{U}/E_A$ and $\mathbb{U}/E_{A\cup\{D\}}$, i.e., the value of

$d_I(\mathbb{U}/E_A, \mathbb{U}/E_{A\cup\{D\}})$ generally decreases. Although this may not be universally true, it is often the case and hence, it is taken as the underlying heuristic to develop the present work. As such, the proposed AR method intuitively employs the smallest value of $d_I(\mathbb{U}/E_A, \mathbb{U}/E_{A\cup\{D\}})$ at each iteration to guide the search for a desirable reduct. In particular, given an attribute

$$r = \text{argmin } d_I(\mathbb{U}/E_{A\cup\{r\}}, \mathbb{U}/E_{A\cup\{r\}\cup\{D\}}), \forall a \in \mathbb{C} - A, \tag{26}$$

if $d_I(\mathbb{U}/E_A, \mathbb{U}/E_{A\cup\{D\}}) > d_I(\mathbb{U}/E_{A\cup\{r\}}, \mathbb{U}/E_{A\cup\{r\}\cup\{D\}})$, the process of reduct search aims to track the greatest inconsistency in a descent manner; else, the process aims to track the lowest inconsistency in an ascent manner.

The second procedure is to determine when to terminate the process of reduct search. As discussed previously, $d_I(\mathbb{U}/E_A, \mathbb{U}/E_{A\cup\{D\}})$ reflects the cost of transforming $\mathbb{U}/E_A$ into $\mathbb{U}/E_{A\cup\{D\}}$. A small value of $d_I(\mathbb{U}/E_A, \mathbb{U}/E_{A\cup\{D\}})$ indicates that the partition distribution of $\mathbb{U}/E_A$ and that of $\mathbb{U}/E_{A\cup\{D\}}$ are similar to each other. According to the equivalence relation defined by Eq. (9), such resemblance implies that $\mathbb{U}/E_A$ and $\mathbb{U}/E_{\{D\}}$ may share many common partitions, and that the condition attribute subset A has a comparable discernibility regarding the decision attribute D. From this point of view, as shown in Eq. (27), the attribute reduct $R \subseteq \mathbb{C}$ is therefore deemed to have an identical discernibility with $\mathbb{C}$, concerning the decision attribute D, namely,

$$d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}}) = d_I(\mathbb{U}/E_{\mathbb{C}}, \mathbb{U}/E_{\mathbb{C}\cup\{D\}}). \tag{27}$$

Thus, the search process can be terminated once this is satisfied. Integrating this procedure with Alg. 1 leads to a robust AR algorithm as presented in Alg. 2.

Note that the time complexity for calculating the agreement matrices for the entire information system is $O(|\mathbb{C}| \times |\mathbb{U}|^2)$. For data denoising, the computation is of the complexity $O(L \times |\mathbb{U}|^2)$, and the time complexity of carrying out the reduction is $O(|\mathbb{C}|^2 \times |\mathbb{U}|^2)$. Therefore, the overall time complexity of Alg. 2 is $O((|\mathbb{C}| + L + |\mathbb{C}|^2) \times |\mathbb{U}|^2)$.

4. Experimental Evaluation

In this section, the effectiveness of the proposed instance denoising and AR methods are experimentally investigated. In particular, Section 4.1 describes the basic set-up for the experimental environment, and Section 4.2

Algorithm 2 Robust attribute reduction

Input:
$I = (\mathbb{U}, \mathbb{C}, D, V, \phi)$, an information system;
k, number of nearest neighbours in Alg. 1.
Output: R, attribute reduct.

1: $\mathbb{U}_0 \leftarrow$ Alg. 1; //Instance denoise.
2: $\gamma \leftarrow d_I(\mathbb{U}_0/E_{\mathbb{C}}, \mathbb{U}_0/E_{\mathbb{C}\cup\{D\}})$
3: $R \leftarrow \emptyset$;
4: **do**
5: $\quad r \leftarrow \text{argmin } d_I(\mathbb{U}_0/E_{A\cup\{r\}}, \mathbb{U}_0/E_{A\cup\{r\}\cup\{D\}}), \forall a \in \mathbb{C} - R$;
6: $\quad R \leftarrow R \cup \{r\}$;
7: **until** $d_I(\mathbb{U}_0/E_R, \mathbb{U}_0/E_{R\cup\{D\}}) = \gamma$
8: **return** R.

examines the effect of using different numbers of the nearest neighbours on the process of denoising data. In Section 4.3, a range of artificial datasets are devised to evaluate the robustness of the proposed approach to detect redundant attributes. Section 4.4 illustrates the changing trend of the value of $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}})$ as the process of AR progresses. Sections 4.5 and 4.6 present a comparative study of the results on reduced datasets, in terms of the returned reduct size and run time, respectively. Sections 4.7 and 4.8 further report on the accuracy and F1-measure of employing the resulting attribute subsets to perform various classification tasks.

4.1. Experimental setup

As shown in Table 3, the experiments are run on 14 datasets taken from the UCI repository of machine learning databases [3] and Knowledge Extraction based on Evolutionary Learning (KEEL) [1]. To facilitate the evaluation, these datasets are discretised using the popular k-means clustering algorithm. Note that, for simplicity, the experiments in Sections 4.2, 4.4 and 4.3 are conducted on the first 9 datasets in Table 3 with 4 clusters in k-means. However, Sections 4.5, 4.6, 4.7 and 4.8 are concerned with the results obtained over all of the datasets in Table 3, with a random number of the clusters in k-means, ranging from 3 to 9 (to minimise any potential adverse impact of using a specific discretisation). Moreover, in order to demonstrate the robustness of the proposed approach, the experiments in the Sections 4.2, 4.3, 4.4 and 4.8 are conducted in the presence of 10% additive random noise to each dataset.

In Sections 4.5, 4.6 and 4.7, apart from the addition of 10% noise to the data, the experimental investigations also consider datasets involving 5% and 15% added noise, to provide a comprehensive evaluation.

Table 3: Evaluation datasets

Dataset	Objects	Attributes	Classes
cleveland	177	13	2
credit	187	15	2
dermatology	366	33	6
forest type	325	27	4
house-vote	435	16	2
ionosphere	351	33	2
promoter	106	58	2
spectfheart	267	44	2
wdbc	569	30	2
colon	62	2000	2
leukemia	72	7129	2
lung	203	3312	5
lymphoma	66	4026	3
segmentation	2100	19	7

Stratified 10×10-fold cross-validation (10×10-FCV) is employed throughout the experimentation. In each 10-FCV, an original dataset is partitioned into 10 subsets of data objects. Of these 10 subsets, a single subset is retained as the testing data for the (subsequent) classifier that uses the original or reduced dataset, and the remaining 9 subsets are used for training. The cross-validation process is repeated for 10 times. The 10 sets of results are then averaged to produce a single estimation of classifier accuracy. The advantage of 10-FCV over random sub-sampling is that all objects are used for both training and testing, and each object is used for testing only once per validation. The stratification of the data prior to its division into folds ensures that each class label (as far as possible) has equal representation in all folds, thereby helping to alleviate bias/variance problems [2].

4.2. Effect of number of nearest neighbours

The impact of k, the number of the nearest neighbours, upon the results of instance denoising is examined here (without using Alg. 2 to reduce the attributes). Particularly, the classification results are obtained using the Logistic and random forest with 10 trees (RF(10)) methods. The average of

the classification results on all of those denoised datasets in relation to the different values of k (ranging from 1 to 20 with a step size of 1) are shown in Fig. 6, via running the two classifiers using the denoised data.

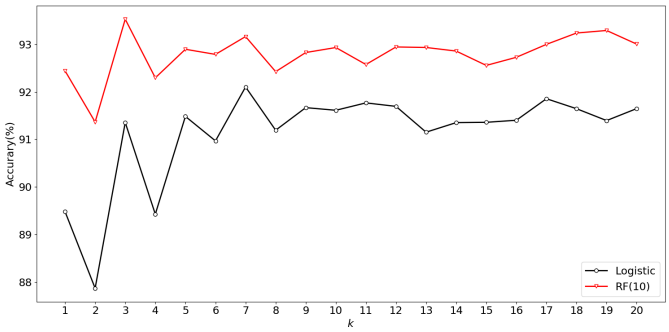
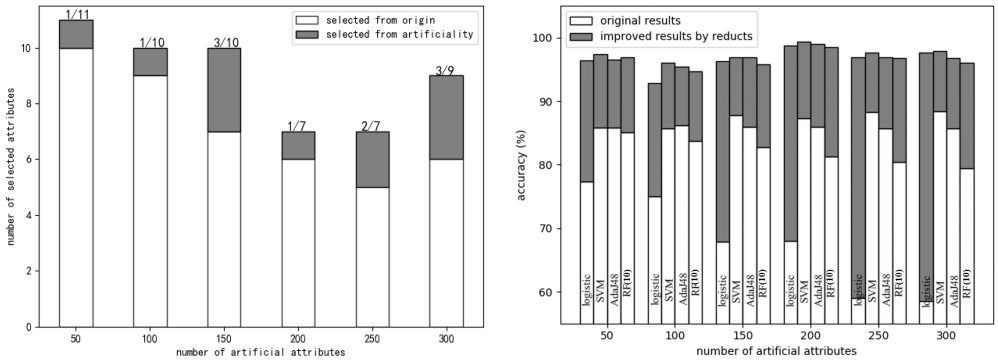


Figure 6: Average classification results

It can be seen from Fig. 6 that with the increase in k, the overall classification accuracies show an upward trend. When the value of k is between 1 and 8, a small neighbourhood of samples leads to less effective noise reduction while revealing a strong volatility. Yet, if the value of k becomes quite large, the samples from minority classes become regarded as noise with high probability, because their proportion in the neighbourhood degrades. Both of these observations are not surprising, meeting the usual expectation that has long been established in the field of electronic engineering. A balance is therefore required between the potential noise smoothing power and the minimisation of the possibility of non-noise data being removed. Recognising this point, while considering that a very small k typically leads to strong fluctuations in classification results and a rather large k tends to result in a time-consuming denoising process, in the following experiments, k is empirically set to 9 (which entails relatively stable classification results as reflected by Fig. 6).

4.3. Impact of denoising strategy

To investigate the effect of applying the proposed denosing method, this set of experiments is carried out on a range of artificially generated datasets. Recall Table 3, the *dermatology* dataset contains the greatest quantity of categories amongst the first 9 datasets listed there. Given its complexity of data classes this dataset is taken to serve as the foundation upon which to generate the artificial datasets. Particularly, six new datasets are created,



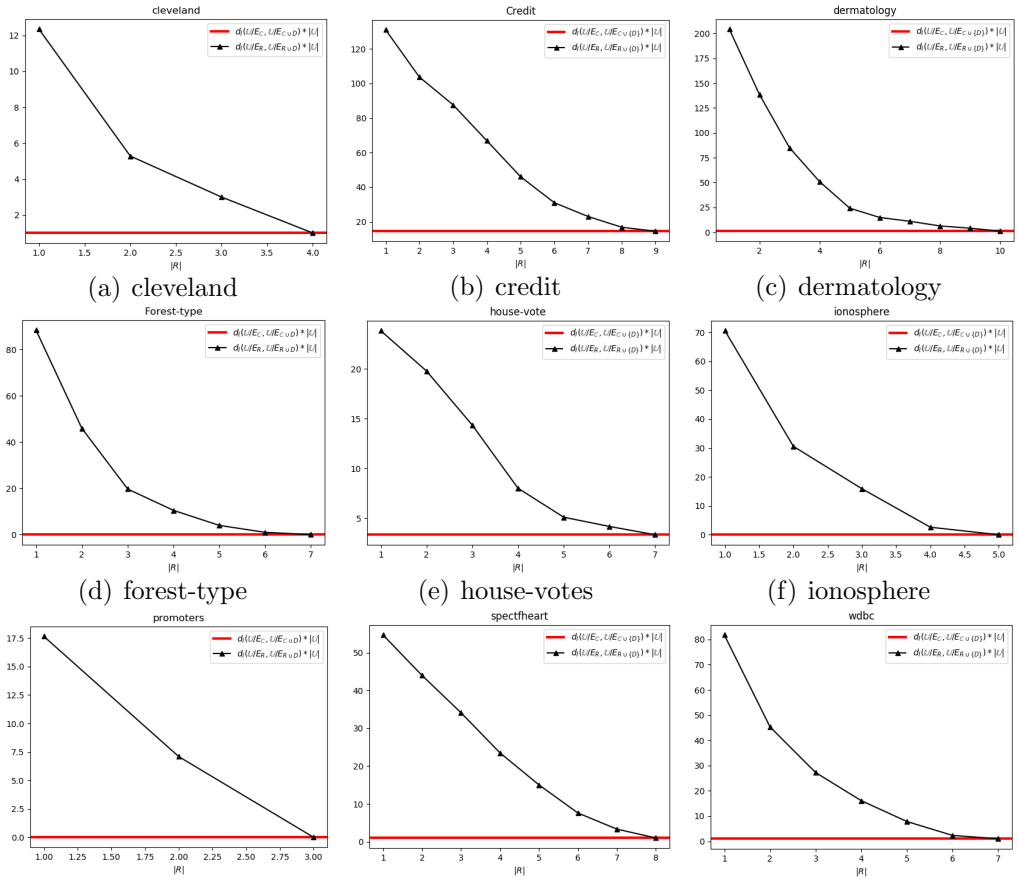
(a) Composition of reduct (b) Classification accuracy

Figure 7: Experimental results on artificial datasets

with each containing 366 instances as per the original *dermatology* dataset but each with 50, 100, 150, 200, 250 or 300 rather low-quality attributes added. To construct the low-quality attributes in each artificial dataset, for every newly added attribute, 90% of its values are set to be of the same value with the remaining 10% assigned randomly. Clearly, this makes each artificial attribute (and henceforth, every generated dataset) to be of rather poor quality, in view of their discriminating power for the classification of the data instances into one of the six different classes. To enable fair comparison, four popular classifiers are employed for this investigation: Logistic, SVM, Adaboost with J48 (AdaJ48), and RF(10).

The composition of the reducts produced by the proposed approach is illustrated in Fig. 7(a). As can be seen, in the returned reducts, most of the attributes are from the original *dermatology* dataset. For example, in the case with 200 or 300 artificial attributes, the reduct only contains 1 or 3 artificial attributes, respectively. This demonstrates that d_I can effectively detect (and hence, remove) low-quality attributes.

Fig. 7(b) shows the comparison results between the use of the reducts (against the number of artificial attributes added) and that of the original dataset on classification accuracy. It can be seen that due to the impact of the low-quality attributes, the classification accuracies on the unreduced datasets are all lower than 90%. However, on the reduced datasets by d_I, all of the accuracies are over 90%. In particular, with the case involving 250 and 300 artificial attributes, the improved accuracies gained by d_I are each over 30%.



(a) cleveland (b) credit (c) dermatology

(d) forest-type (e) house-votes (f) ionosphere

(g) promoters (h) spectfheart (i) wdbc

Figure 8: Inconsistency measure with respect to d_I and size of reduct $|R|$

4.4. Changes of inconsistency measure during AR process

Fig. 8 shows the results of measuring the inconsistency with the proposed AR approach across all datasets studied, where the red lines represent the baseline $d_I(\mathbb{U}/E_{\mathbb{C}}, \mathbb{U}/E_{\mathbb{C}\cup\{D\}})$ and the black lines stand for $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}})$. Note that rather than $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}})$, $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}}) \cdot |\mathbb{U}|$ for each dataset is illustrated in relation to the growth of reduct size $|R|$, in an effort to highlight the trend of inconsistency measure.

It can be observed that $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}}) \cdot |\mathbb{U}|$ is monotonically decreasing as AR progresses until it equals to $d_I(\mathbb{U}/E_{\mathbb{C}}, \mathbb{U}/E_{\mathbb{C}\cup\{D\}}) \cdot |\mathbb{U}|$. These experimental results clearly demonstrate the effectiveness of the proposed AR mechanism. Moreover, the observed minuscule values of $d_I(\mathbb{U}/E_{\mathbb{C}}, \mathbb{U}/E_{\mathbb{C}\cup\{D\}})$

(or those of $d_I(\mathbb{U}/E_R, \mathbb{U}/E_{R\cup\{D\}})$) imply that the distribution of the partitions of a given dataset does not change much after being refined with respect to the decision attribute. That is, the corresponding dataset has an entire (or reduced) set of condition attributes which are highly consistent to the decision attribute.

4.5. Comparison on reduct size

By summarising the size of the reducts over each dataset as given in Fig. 8, a comparison is herein made on the reduced dataset size between the proposed (namely, d_I-based) approach and those achieved using either of the seven existing AR methods: ASNAR [13], FKD [35], GBNRS [33], RelieF [31], VPRS [32], and SPDTRS [29]. The configuration of these methods are shown in Table 4. Note that, since RelieF filters attributes by rank and selects top scoring ones, the reduct size of each dataset returned by RelieF is set to be identical to that by d_I in the following experiments.

Table 4: Configuration of alternative AR methods compared

Methods	Parameters
ASNAR	$\delta_O = 0.9 \times \delta_I$, $\delta_I = 0.3$
FBD	$\delta(\widetilde{A}, \widetilde{B}) = \frac{\sum_{x\in U} \mu_{\widetilde{A}\cup\widetilde{B}}(x) - \sum_{x\in U} \mu_{\widetilde{A}\cap\widetilde{B}}(x)}{\lvert U\rvert}$
GBNRS	purity=1
RelieF	$k = \max\{\lvert\mathbb{U}\rvert, 100\}$
VPRS	$\beta = 0.3$
SPDTRS	ζ=0.2

The results in Table 5 collectively show that, with any of the three levels of added noise, the reduct size obtained by the d_I-based approach is smaller than those attainable with either of the alternative methods, on most of the 14 datasets. For example, on the datasets *promoters*, *colon* and *leukemia*, d_I-based selects only 3, 2 and 2 attributes to form the returned reduct, respectively, whilst the reducts returned by the alternatives are much larger. Particularly, across all three noise levels, the average size over all 14 datasets is much smaller than that achievable by any other method compared. Such strong performance may be attributed to the data denoising procedure that reduces the inconsistency in the datasets. The following section further investigates the classification performance of using selected attribute subsets, showing that the returned reducts by the d_I-based AR mechanism also retains sufficient information to entail high discriminating ability.

Table 5: Reduct sizes

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	5%	3	6	7	4	3	11	4
	10%	3	7	8	5	3	11	4
	15%	5	7	7	5	5	12	7
credit	5%	8	14	10	9	8	11	11
	10%	8	14	10	10	8	15	13
	15%	11	14	10	10	11	14	13
dermatology	5%	10	11	12	18	10	27	7
	10%	10	10	11	13	10	30	7
	15%	10	10	11	12	10	31	8
forest type	5%	6	8	5	5	6	11	5
	10%	5	9	5	5	5	9	5
	15%	6	9	5	4	6	10	5
house-votes	5%	7	12	10	10	7	12	10
	10%	5	13	11	9	5	16	12
	15%	12	14	12	11	12	16	12
ionosphere	5%	5	7	7	11	5	5	5
	10%	5	8	9	10	5	30	7
	15%	5	11	10	10	5	28	7
promoters	5%	3	5	4	7	3	10	5
	10%	3	5	4	7	3	4	5
	15%	4	5	4	6	4	4	6
spectfheart	5%	5	7	5	10	5	7	5
	10%	5	7	5	8	5	7	5
	15%	5	7	5	9	5	7	5
wdbc	5%	5	9	5	15	5	16	5
	10%	4	9	5	10	4	22	5
	15%	5	9	5	10	5	23	6
colon	5%	2	4	4	5	2	2	3
	10%	2	4	4	5	2	457	3
	15%	2	4	4	2	2	479	3
leukemia	5%	2	4	4	5	2	822	3
	10%	2	3	5	3	2	1030	3
	15%	2	4	5	4	2	2209	3
lung	5%	4	5	5	6	4	286	4
	10%	3	6	5	6	3	79	3
	15%	3	6	4	5	3	1004	4
lymphoma	5%	2	3	4	2	2	451	2
	10%	2	3	5	3	2	423	3
	15%	2	4	4	2	2	904	3
segmentation	5%	19	15	11	16	19	19	15
	10%	13	15	13	15	13	19	15
	15%	19	15	15	16	19	19	15
average	5%	5.8	7.9	6.6	8.8	5.8	120.7	6.0
	10%	5.0	8.1	7.1	7.8	5.0	153.4	6.4
	15%	6.5	8.5	6.8	7.6	6.5	340.0	6.9

4.6. Evaluation on runtime cost

In addition to the experimental results on the reduct size, the runtime consumed to conduct AR is also displayed in Table 6. Note that most algorithms in this paper are programmed in Python, but RelieF is executed via a software package [11] and SPDTRS is coded in Matlab. Since the platforms to implement these AR methods are different, the runtime costs incurred by them are not directly comparable without a unified standard. Nonetheless, it can be seen that, for the wide range of the datasets employed in the experimental investigations, the time spent to calculate d_I is practically acceptable (being around 10 minutes in the worst case) in general.

4.7. Comparison on classification accuracy

This set of experiments is carried out to provide a systematic comparison regarding the classification accuracy on the reduced datasets, again amongst the following methods: d_I-based, ASNAR, VPRS, FKD, GBNRS, RelieF and SPDTRS. This study is also performed in conjunction with the use of Logistic, SVM, AdaJ48 or RF(10). The respective results are shown in Tables 7, 8, 9 and 10, where the average classification accuracies gained using 10-FCV for each of the methods are recorded, with the best results for each dataset underlined. In addition, the number of the best performances attained by each AR method is summarised in the bottom row within each of these tables.

It can be seen that in conjunction with the use of either Logistic, SVM, AdaJ48, or RF(10), across all different noise level setings (5%/10%/15%), the classification performance of the proposed method is superior to those attainable by the existing methods on 5/8/5, 3/5/7, 4/8/3 or 7/7/5 out of the 14 datasets. Occasionally, with 15% added noise, RelieF generates the best average accuracy and most of such best results are linked with the use of Logistic and AdaJ48. Importantly, the overall excellent performance of the d_I-based method is on average, achieved through the use of the smallest attribute subsets returned by it. For those datasets where the use of d_I-returned attributes does not lead to the highest accuracy, the performances remain compatible to the rest, but mostly involving far less attributes. Moreover, the use of the proposed approach does not lead to the poorest performance in most cases, resulting in an accuracy generally well above the average instead.

Table 6: Results of runtime (s)

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	5%	1.8581	6.1623	0.1237	107.3635	0.1257	1.9743	1.8735
	10%	1.9817	6.6054	0.1496	113.0802	0.1297	2.1612	1.9471
	15%	1.9631	6.7891	0.1397	137.6155	0.1416	2.3815	2.0414
credit	5%	34.5424	88.7223	1.1290	106.2250	0.4787	37.0736	11.9832
	10%	36.5814	89.0803	1.2233	183.6076	0.5306	38.3715	13.1349
	15%	39.3947	108.8365	1.2995	263.3471	0.5795	42.7083	13.9048
dermatology	5%	21.3398	72.2145	0.8318	143.9442	0.2872	21.3519	13.3359
	10%	22.0933	88.0192	0.9380	431.8071	0.2952	23.3616	15.3019
	15%	22.4163	89.7175	0.9853	443.7204	0.3241	25.6255	17.0929
forest type	5%	10.7514	41.0641	0.5685	115.5908	0.2414	13.4579	9.4535
	10%	11.2143	55.5115	0.6304	250.2127	0.2563	14.7732	11.4191
	15%	11.5003	69.9348	0.6662	347.7581	0.2803	16.1233	11.9097
house-votes	5%	14.1368	40.3852	0.5475	226.4088	0.3032	14.6051	8.3987
	10%	15.1608	42.4943	0.6413	147.0161	0.3281	15.9749	9.0111
	15%	16.4146	52.6708	0.6562	183.3275	0.3521	17.7673	9.2401
ionosphere	5%	16.4975	61.6112	0.8796	143.5376	0.2673	19.4981	15.4189
	10%	17.8382	96.3812	0.9923	591.6166	0.2882	21.3827	21.0503
	15%	18.2362	99.4010	1.0492	341.5810	0.3072	23.4971	23.0135
promoters	5%	2.3483	12.2303	0.2922	98.5833	0.0838	3.0229	4.2616
	10%	2.2650	13.0112	0.3281	145.2164	0.0888	3.2889	4.4805
	15%	2.6268	16.1636	0.3341	259.4777	0.0967	3.7011	4.6498
spectfheart	5%	12.7280	46.9105	0.8348	131.4935	0.2284	14.5194	12.2664
	10%	12.5938	50.7193	0.9295	346.2840	0.2284	16.3795	14.0960
	15%	12.3711	58.8308	0.9525	487.2099	0.2773	18.2677	13.6461
wdbc	5%	44.8893	92.9480	1.7473	126.0470	0.4428	46.7439	22.9993
	10%	48.4196	122.0870	1.9398	392.0205	0.4757	51.7120	25.8229
	15%	50.6763	136.5664	2.0446	263.7432	0.5057	57.0793	27.5080
colon	5%	22.5079	93.2303	44.7551	1482.2134	0.4198	35.2612	40.9748
	10%	23.1776	98.5275	44.1459	3668.5586	0.4098	144.6029	43.3100
	15%	26.4958	102.8152	46.6140	5982.7483	0.4498	158.1656	44.8791
leukemia	5%	133.4688	379.2381	630.9751	10621.9400	1.3393	1116.4087	171.1494
	10%	142.5839	317.8804	646.8009	10803.9733	1.4292	1351.2185	179.7040
	15%	146.8416	439.6555	674.9973	18241.5824	1.4792	2550.8980	189.0943
lung	5%	603.2815	878.0916	566.4605	23936.2678	2.4387	1218.0571	333.4656
	10%	577.0184	1119.6728	636.6002	26256.8426	2.4487	831.2887	294.1929
	15%	559.3375	1182.3693	633.9602	28635.1598	2.6286	2997.3574	371.4187
lymphoma	5%	81.5365	152.5186	172.7093	4983.5417	0.6696	272.9826	69.0056
	10%	86.6313	160.5943	183.6443	9976.1314	0.7596	281.7306	90.5938
	15%	87.9231	220.3025	189.3460	11878.773	0.7296	525.2047	96.8955
segmentation	5%	410.6343	991.3292	7.7059	460.1845	2.4387	438.6860	51.9757
	10%	430.8801	1112.8319	8.2256	727.6519	2.5386	483.9518	53.3017
	15%	436.5172	1265.5772	8.6253	853.9045	2.5086	530.2072	58.8428
average	5%	100.7515	211.1897	102.1115	3169.3175	0.6975	232.4031	54.7544
	10%	102.0314	240.9583	109.0849	3859.5728	0.7291	234.2998	55.5262
	15%	102.3368	274.9736	111.5479	4879.9963	0.7615	497.7846	63.1526

Table 7: Classification accuracy by Logistic (%)

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	5%	87.35	86.17	86.92	88.06	89.08	85.42	87.25
	10%	85.24	81.41	80.99	86.28	85.19	81.55	85.06
	15%	82.35	78.46	79.50	80.65	83.56	78.91	82.38
credit	5%	84.54	82.94	83.86	83.03	84.11	83.40	75.83
	10%	83.24	80.46	81.55	80.20	82.77	80.55	73.58
	15%	78.15	77.19	78.22	77.61	78.63	77.57	78.41
dermatology	5%	87.67	77.3	68.8	89.34	87.07	88.35	66.68
	10%	70.50	70.93	58.9	72.42	80.10	69.3	66.56
	15%	70.34	62.13	67.53	67.76	75.05	65.58	67.90
forest type	5%	73.25	69.67	32.28	62.69	69.65	69.53	69.69
	10%	71.46	66.22	33.38	60.85	63.92	73.02	67.45
	15%	68.23	64.34	33.53	42.29	66.32	67.28	62.19
house-votes	5%	94.51	93.35	88.93	93.94	93.79	93.00	94.12
	10%	90.69	90.42	91.11	89.84	90.63	90.04	90.29
	15%	89.10	84.26	85.04	89.14	88.66	88.46	89.52
ionosphere	5%	86.60	88.35	81.16	81.44	87.24	88.87	87.36
	10%	85.75	82.44	79.41	79.61	83.14	67.52	82.40
	15%	85.30	82.09	79.3	79.16	82.05	67.01	82.84
promoters	5%	87.45	76.36	56.09	60.91	87.82	70.18	83.82
	10%	85.80	85.60	55.58	57.89	82.11	79.35	64.60
	15%	81.67	78.67	55.58	55.67	82.67	77.33	74.67
spectfheart	5%	74.09	73.11	73.09	68.21	76.03	72.58	76.92
	10%	72.24	70.84	72.67	71.32	71.68	71.34	73.29
	15%	69.30	67.03	69.74	64.84	71.26	68.23	69.45
wdbc	5%	92.99	89.95	88.41	88.59	89.45	91.14	90.15
	10%	90.91	88.64	86.56	88.74	88.14	85.28	84.40
	15%	89.27	86.54	85.07	86.98	86.02	83.58	88.18
colon	5%	75.02	67.93	47.19	59.86	75.98	80.69	64.05
	10%	75.62	61.67	48.62	51.74	70.98	56.55	61.79
	15%	72.80	64.84	47.45	63.96	69.21	60.07	46.18
leukemia	5%	89.18	84.02	50.59	49.36	90.16	88.59	66.21
	10%	89.75	92.23	47.54	45.68	91.20	87.91	65.23
	15%	84.44	84.04	52.43	53.08	87.93	79.99	58.44
lung	5%	77.11	77.44	67.45	59.48	83.11	83.58	60.96
	10%	84.02	72.16	65.05	59.98	82.40	68.37	68.00
	15%	74.1	67.98	64.29	60.30	77.62	62.06	58.70
lymphoma	5%	86.38	80.60	50.81	53.93	85.67	86.10	82.07
	10%	87.93	82.79	46.50	55.89	79.27	80.34	65.84
	15%	76.75	79.93	51.27	57.93	75.71	82.96	75.57
segmentation	5%	87.73	87.82	81.07	87.93	87.73	87.73	87.82
	10%	83.65	83.31	82.64	83.39	83.27	83.04	83.31
	15%	80.49	80.75	80.52	80.49	80.49	80.49	80.75
average	5%	84.56	81.07	68.33	73.34	84.78	83.51	78.07
	10%	82.63	79.22	66.46	70.27	81.06	76.73	73.70
	15%	78.74	75.59	66.39	68.56	78.94	74.25	72.51
best	5%	5	0	0	1	4	2	2
	10%	8	1	1	1	1	1	1
	15%	5	1	0	0	7	1	1

Table 8: Classification accuracy by SVM (%)

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	5%	88.87	87.95	88.76	87.36	89.04	88.76	88.65
	10%	83.39	86.22	85.88	83.13	84.21	86.95	85.05
	15%	81.94	79.46	81.14	82.03	84.11	82.78	83.37
credit	5%	84.21	83.72	83.80	83.83	83.94	84.00	75.14
	10%	82.65	82.60	82.23	81.73	82.43	82.62	73.46
	15%	78.78	78.48	78.76	78.96	79.46	78.85	79.14
dermatology	5%	93.33	87.51	72.07	92.83	91.13	93.98	70.53
	10%	87.81	82.82	64.11	88.28	85.83	85.61	71.47
	15%	87.11	66.20	77.37	85.22	78.72	80.27	71.48
forest type	5%	76.29	76.43	34.49	66.34	69.59	78.98	71.16
	10%	72.58	69.66	34.35	64.53	66.09	75.67	67.92
	15%	71.62	67.50	33.45	44.80	69.04	73.67	66.01
house-votes	5%	94.10	93.90	89.11	94.06	94.23	94.10	94.01
	10%	90.99	90.78	90.41	90.78	90.92	90.61	90.71
	15%	90.14	85.28	84.82	90.16	90.18	90.06	90.16
ionosphere	5%	87.07	90.33	82.09	88.45	87.51	88.87	86.65
	10%	85.72	85.46	80.39	81.28	83.64	80.59	83.64
	15%	86.70	86.54	79.72	79.04	83.34	82.76	82.92
promoters	5%	84.73	76.82	56.36	63.45	88.82	75.09	85.73
	10%	83.23	85.20	54.80	59.77	84.28	76.94	73.83
	15%	81.50	80.25	51.17	57.08	83.75	77.50	73.50
spectfheart	5%	76.05	73.87	75.31	71.11	76.71	76.24	75.56
	10%	72.33	70.60	72.87	73.64	74.07	72.29	71.28
	15%	69.94	70.31	69.16	66.31	71.91	70.37	71.91
wdbc	5%	93.01	91.18	88.04	92.51	89.50	92.25	88.98
	10%	90.91	89.47	86.04	91.15	87.68	88.03	85.01
	15%	88.31	88.71	85.70	88.27	85.90	88.52	87.29
colon	5%	70.29	59.71	48.40	52.90	67.48	81.05	58.24
	10%	75.86	67.90	47.86	58.88	67.48	72.33	62.24
	15%	71.68	64.84	48.40	61.11	67.48	66.8	48.43
leukemia	5%	90.91	87.52	56.61	54.14	90.41	91.45	64.52
	10%	91.46	89.68	53.73	48.32	88.73	91.73	64.11
	15%	89.03	87.83	55.32	55.10	87.93	84.01	57.99
lung	5%	87.53	81.47	68.49	70.95	83.07	86.43	74.50
	10%	84.84	77.01	66.64	66.59	81.77	80.31	70.81
	15%	77.19	76.76	64.97	65.76	76.44	74.02	65.84
lymphoma	5%	91.17	85.17	65.29	66.00	87.29	94.05	83.88
	10%	88.71	90.96	53.57	60.55	78.18	85.71	76.68
	15%	86.59	85.00	60.87	63.04	76.82	83.71	81.96
segmentation	5%	88.83	88.78	81.84	88.77	88.83	87.18	88.78
	10%	84.41	84.68	83.19	84.68	84.00	84.77	84.68
	15%	81.64	81.57	81.29	81.37	81.64	81.64	81.57
average	5%	86.10	83.17	70.76	76.62	84.83	86.72	79.02
	10%	83.92	82.36	68.31	73.81	81.10	82.44	75.78
	15%	81.58	78.48	67.95	71.30	79.38	79.63	74.40
best	5%	3	1	0	0	5	5	0
	10%	5	2	0	2	1	4	0
	15%	7	1	0	0	6	2	1

Table 9: Classification accuracy by AdaJ48 (%)

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	5%	86.93	83.91	88.40	87.52	87.68	85.37	88.16
	10%	84.67	77.72	77.07	84.95	83.70	80.61	83.56
	15%	75.89	74.24	72.96	78.73	81.49	74.03	77.58
credit	5%	79.19	79.62	79.86	78.84	83.00	79.64	70.76
	10%	77.42	77.28	76.73	75.94	81.40	77.06	67.49
	15%	71.37	71.30	71.76	71.64	75.79	71.40	72.41
dermatology	5%	89.82	84.91	71.10	91.36	87.00	92.49	67.15
	10%	82.89	78.53	62.50	84.44	82.81	83.04	69.40
	15%	82.32	59.33	73.95	80.43	74.62	79.93	66.63
forest type	5%	73.52	72.16	40.51	61.92	69.63	73.66	70.63
	10%	70.94	67.00	37.74	62.29	67.60	75.11	66.15
	15%	65.26	60.89	39.60	43.23	67.57	70.27	64.16
house-votes	5%	93.75	92.61	85.93	93.12	92.44	91.82	92.56
	10%	91.70	86.44	88.43	87.15	89.98	86.40	87.99
	15%	85.70	79.46	79.58	87.20	86.66	85.40	86.48
ionosphere	5%	87.88	88.01	83.3	88.42	86.77	91.27	86.80
	10%	83.93	80.30	77.38	77.96	81.72	80.83	81.23
	15%	83.85	82.44	77.81	79.33	81.00	81.89	81.60
promoters	5%	89.45	81.18	56.91	61.00	86.91	75.09	81.45
	10%	87.86	81.43	56.48	56.58	83.55	76.33	62.33
	15%	81.58	77.33	55.42	57.75	82.17	75.83	72.92
spectfheart	5%	71.91	71.29	70.98	68.99	73.79	71.40	73.32
	10%	71.77	71.64	72.45	71.81	68.35	68.97	74.15
	15%	67.43	65.03	68.59	64.59	68.01	64.81	65.87
wdbc	5%	91.29	89.84	86.07	90.57	89.26	90.86	88.07
	10%	88.68	84.99	81.89	86.66	87.36	86.53	83.97
	15%	86.42	81.36	82.21	84.00	84.86	83.72	84.37
colon	5%	78.90	80.29	57.57	60.62	73.71	80.71	61.05
	10%	71.19	65.48	54.60	52.98	69.31	64.48	64.12
	15%	67.48	72.98	53.02	69.09	65.89	62.46	48.07
leukemia	5%	91.16	90.14	54.25	54.32	92.00	87.75	60.71
	10%	91.43	95.13	51.52	48.79	89.14	85.36	60.54
	15%	87.40	87.29	56.03	61.53	89.76	80.33	55.17
lung	5%	84.65	78.63	63.81	61.32	81.45	80.99	69.07
	10%	84.16	75.47	62.29	59.95	79.24	75.34	69.00
	15%	70.42	69.09	61.29	56.97	74.20	69.34	59.69
lymphoma	5%	87.1	88.43	61.83	67.62	89.67	87.71	84.88
	10%	84.48	83.73	58.93	64.25	76.20	79.75	76.79
	15%	79.57	82.18	65.11	65.11	80.25	78.05	77.64
segmentation	5%	88.83	88.98	83.39	88.78	88.83	88.83	88.98
	10%	84.4	84.17	83.79	84.10	84.21	84.45	84.17
	15%	81.65	81.35	81.14	81.77	81.65	81.65	81.35
average	5%	85.31	83.57	70.28	75.31	84.44	84.11	77.40
	10%	82.54	79.24	67.27	71.28	80.33	78.88	73.64
	15%	77.60	74.59	67.03	70.10	78.14	75.65	71.00
best	5%	4	1	0	1	5	3	1
	10%	8	1	0	2	1	1	1
	15%	3	2	1	1	6	1	0

Table 10: Classification accuracy by RF(10)(%)

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	F-score
cleveland	5%	86.44	87.04	88.10	87.95	88.44	86.67	87.36
	10%	83.59	77.94	77.94	84.1	84.98	79.32	81.61
	15%	75.16	72.35	72.55	75.57	80.70	73.58	73.89
credit	5%	80.46	80.88	80.51	80.18	82.74	80.78	70.14
	10%	77.52	77.94	77.53	76.71	80.15	78.21	67.53
	15%	71.26	71.89	71.1	70.83	74.57	72.03	72.68
dermatology	5%	89.27	84.89	71.15	90.79	88.25	91.05	67.78
	10%	81.85	77.64	62.69	83.44	81.72	84.16	68.88
	15%	80.55	59.66	70.51	78.73	73.42	80.31	63.56
forest type	5%	73.64	72.95	35.87	60.75	69.16	74.52	67.56
	10%	69.17	67.35	34.72	59.54	67.94	74.31	63.92
	15%	64.18	61.99	34.12	37.28	62.31	67.28	58.60
house-votes	5%	93.86	93.16	86.54	93.38	92.79	92.28	92.96
	10%	91.57	87.09	88.85	87.18	89.84	87.46	88.49
	15%	86.56	80.86	80.38	86.82	87.08	86.58	86.90
ionosphere	5%	88.23	88.69	83.69	87.93	86.44	90.84	86.03
	10%	84.01	81.05	78.68	80.06	81.78	82.42	81.26
	15%	83.23	82.72	78.98	79.58	80.30	82.64	81.30
promoters	5%	89.09	82.09	53.55	61.27	85.45	74.36	81.45
	10%	87.11	81.28	55.58	56.30	82.25	76.30	60.51
	15%	79.17	76.08	52.58	53.00	78.42	71.25	72.50
spectfheart	5%	74.81	73.27	69.87	71.47	70.54	71.11	71.9
	10%	71.33	72.19	69.86	74.15	67.67	69.6	73.84
	15%	65.59	65.22	64.17	66.71	63.37	64.71	62.92
wdbc	5%	91.62	89.99	86.39	91.11	89.43	92.13	88.34
	10%	88.47	86.32	81.76	87.97	87.57	87.42	83.49
	15%	86.45	83.48	80.63	85.21	85.09	85.42	84.46
colon	5%	76.74	77.02	54.52	57.86	72.10	83.43	61.98
	10%	73.48	63.67	48.86	51.38	70.90	60.24	56.02
	15%	67.75	65.37	47.27	67.45	67.86	59.25	56.02
leukemia	5%	91.70	86.36	53.93	57.46	90.68	78.09	61.00
	10%	92.36	94.88	54.95	48.96	90.25	80.80	61.54
	15%	88.88	88.18	53.82	61.38	88.67	72.31	54.17
lung	5%	85.09	79.90	63.37	66.49	82.27	79.90	65.65
	10%	84.34	74.92	60.66	60.63	80.95	66.96	65.78
	15%	69.69	69.39	58.92	55.72	77.71	67.06	52.73
lymphoma	5%	89.55	89.33	57.21	58.79	87.50	85.38	84.29
	10%	85.73	84.45	49.68	56.05	77.34	80.91	69.21
	15%	80.88	82.82	50.59	49.09	75.52	77.79	66.93
segmentation	5%	88.14	88.05	81.63	88.06	88.14	88.14	88.05
	10%	82.77	82.56	83.53	82.49	82.70	82.42	82.56
	15%	79.02	78.98	79.53	79.29	79.02	79.02	78.98
average	5%	85.62	83.83	69.02	75.25	83.85	83.48	76.75
	10%	82.38	79.23	66.09	70.64	80.43	77.90	71.76
	15%	77.03	74.21	63.94	67.62	76.72	74.23	68.97
best	5%	7	0	0	0	3	6	0
	10%	7	1	1	1	2	2	0
	15%	5	1	0	2	5	1	0

4.8. Comparison on F1 measure

In this experiment, F1 measure is used to evaluate the performance of the d_I-based method on problems involving class imbalance. The results given in Table 11 demonstrate that the proposed approach is ranked first in terms of the overall count of the best values of F1 measure. In addition, the use of this approach does not lead to the poorest performance in most case, except for the *cleveland* dataset that is classified with SVM.

Together, all of the above results illustrate that the present work entails an overall stronger performance in terms of robustness, reduct size, classification accuracy and F1 measure. Although, occasionally, the use of d_I-based returns lower classification accuracies than the use of RelieF, its overall superiority in F1 measure shows the ability of d_I in relieving data imbalance. In particular, compared to FKD, another EMD-based AR method, the outstanding performance of the present work demonstrates the benefit of utilising the proposed denoising strategy.

5. Conclusion

This paper has presented an EMD-based inconsistency measure to help evaluate the discernibility of an attribute subset with respect to the decision attribute. Particularly, in order to enhance the robustness of attribute reduction, the work utilises a denoising strategy to detect noisy instances. The effectiveness of the proposed instance denoising and AR procedures has been verified with systematic experiments, in the context of being utilised to support performing classification tasks, via testing against popular, state-of-the-art AR methods. Comparative results have demonstrated in general that the proposed AR approach can detect attribute subsets of much smaller in size than state-of-the-art methods, while leading to the achievement of a higher classification accuracy and F1 measure. The runtime cost to produce the reducts following this approach has been shown to be practically reasonable. Overall, the proposed AR approach can effectively overcome the adverse impact caused by noisy data and redundant attributes.

Whilst promising, the work also opens up an interesting avenue for further development. For instance, it would be useful to investigate how it may be extended to handling more complicated large-scale datasets [26], e.g. those requiring higher-order [21, 23] or ensemble [10] classification, or involving multi-label [24] and unsupervised [4, 15] learning. Also, in the present work, datasets are discretised using the popular k-means clustering algorithm. No

Table 11: Results of F1 measure

Datasets	Ratio	d_I	ASNAR	FKD	GBNRS	RelieF	VPRS	SPDTRS
cleveland	Logistic	0.92	0.89	0.89	0.92	0.89	0.92	0.92
	SVM	0.91	0.92	0.92	0.91	0.91	0.93	0.92
	AdaJ48	0.91	0.87	0.91	0.86	0.91	0.89	0.91
	RF	0.91	0.87	0.87	0.91	0.92	0.88	0.90
credit	Logistic	0.83	0.79	0.80	0.79	0.82	0.79	0.69
	SVM	0.83	0.83	0.82	0.82	0.83	0.83	0.66
	AdaJ48	0.75	0.75	0.74	0.73	0.80	0.75	0.64
	RF	0.76	0.76	0.76	0.75	0.79	0.76	0.63
dermatology	Logistic	0.80	0.82	0.89	0.81	0.90	0.81	0.88
	SVM	0.92	0.92	0.94	0.93	0.94	0.92	0.91
	AdaJ48	0.90	0.90	0.91	0.92	0.92	0.91	0.90
	RF	0.91	0.91	0.93	0.93	0.92	0.93	0.89
forest type	Logistic	0.69	0.66	0.28	0.63	0.66	0.73	0.70
	SVM	0.71	0.68	0.28	0.70	0.70	0.75	0.73
	AdaJ48	0.68	0.66	0.25	0.66	0.71	0.74	0.68
	RF	0.65	0.64	0.37	0.62	0.71	0.73	0.65
house-votes	Logistic	0.88	0.88	0.89	0.88	0.88	0.88	0.88
	SVM	0.89	0.89	0.89	0.89	0.89	0.88	0.89
	AdaJ48	0.90	0.83	0.86	0.84	0.88	0.83	0.85
	RF	0.89	0.84	0.86	0.84	0.87	0.85	0.86
ionosphere	Logistic	0.89	0.86	0.84	0.84	0.87	0.73	0.86
	SVM	0.89	0.89	0.85	0.85	0.88	0.84	0.87
	AdaJ48	0.87	0.84	0.82	0.82	0.86	0.85	0.85
	RF	0.87	0.85	0.83	0.84	0.86	0.86	0.85
promoters	Logistic	0.87	0.86	0.54	0.56	0.82	0.80	0.65
	SVM	0.84	0.86	0.55	0.58	0.85	0.77	0.74
	AdaJ48	0.90	0.80	0.57	0.59	0.87	0.73	0.81
	RF	0.90	0.81	0.54	0.59	0.86	0.73	0.81
spectfheart	Logistic	0.36	0.34	0.43	0.44	0.05	0.37	0.36
	SVM	0.29	0.23	0.34	0.44	0	0.31	0.09
	AdaJ48	0.41	0.44	0.42	0.45	0.11	0.38	0.48
	RF	0.39	0.41	0.40	0.46	0.19	0.34	0.49
wdbc	Logistic	0.93	0.91	0.89	0.91	0.91	0.88	0.87
	SVM	0.93	0.92	0.89	0.93	0.90	0.90	0.87
	AdaJ48	0.91	0.88	0.85	0.89	0.90	0.89	0.87
	RF	0.91	0.89	0.85	0.90	0.90	0.90	0.87
colon	Logistic	0.66	0.50	0.22	0.36	0.56	0.33	0.44
	SVM	0.63	0.55	0.14	0.64	0.51	0.55	0.43
	AdaJ48	0.58	0.52	0.11	0.30	0.54	0.50	0.47
	RF	0.61	0.51	0.33	0.28	0.56	0.39	0.38
leukemia	Logistic	0.93	0.94	0.57	0.57	0.93	0.90	0.73
	SVM	0.93	0.92	0.65	0.62	0.91	0.91	0.72
	AdaJ48	0.93	0.96	0.61	0.61	0.92	0.88	0.71
	RF	0.94	0.96	0.66	0.58	0.93	0.86	0.70
lung	Logistic	0.90	0.81	0.78	0.74	0.89	0.78	0.79
	SVM	0.90	0.84	0.79	0.79	0.88	0.85	0.81
	AdaJ48	0.89	0.52	0.74	0.73	0.86	0.83	0.80
	RF	0.89	0.82	0.72	0.73	0.88	0.83	0.78
lymphoma	Logistic	0.65	0.86	0.11	0	0.36	0.87	0.21
	SVM	0.66	0.90	0.02	0	0.36	0.70	0.12
	AdaJ48	0.60	0.90	0.01	0	0.26	0.57	0.16
	RF	0.65	0.90	0	0	0.35	0.47	0.16
segmentation	Logistic	0.91	0.91	0.91	0.91	0.91	0.91	0.92
	SVM	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	AdaJ48	0.90	0.90	0.91	0.90	0.91	0.90	0.90
	RF	0.88	0.88	0.91	0.88	0.89	0.87	0.88
best		29	10	7	8	14	8	7

optimisation is carried out in clustering, simply in order to provide a fair ground upon which to compare different methods (as certain techniques may benefit more from any optimisation than others). Nevertheless, it may be of interest to investigate exactly how a discretisation method may affect the proposed AR algorithm. Moreover, the proposed inconsistency degree d_I may be constructed using one of the many alternative distance measures. Thus, how to efficiently integrate an innovative form of distance measure remains another piece of interesting research.

Acknowledgments

This work is jointly supported by the Innovation Support Plan for Dalian High-level Talents (No. 2018RQ70) and a Sêr Cymru II COFUND Fellowship, UK. The authors are grateful to the anonymous reviewers for their constructive comments, which have helped improve this work significantly.

References

[1] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework., Journal of Multiple-Valued Logic & Soft Computing 17 (2011).

[2] Y. Bengio, Y. Grandvalet, Bias in estimating the variance of K-fold cross-validation, in: Statistical Modeling and Analysis for Complex Data Problems, Springer, 2005, pp. 75–95.

[3] C. Blake, C. Merz, UCI repository of machine learning databases, 1998. University of California, Irvine, School of Information and Computer Sciences.

[4] T. Boongoen, C. Shang, N. Iam-On, Q. Shen, Extending data reliability measure to a filter approach for soft subspace clustering, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41 (2011) 1705–1714.

[5] Y. Chen, K. Liu, J. Song, H. Fujita, X. Yang, Y. Qian, Attribute group for attribute reduction, Information Sciences 535 (2020) 64–80.

[6] J. Dai, H. Hu, W. Wu, Y. Qian, D. Huang, Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets, IEEE Transactions on Fuzzy Systems 26 (2018) 2174–2187.

[7] J. Dai, Q. Hu, H. Hu, D. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, IEEE Transactions on Fuzzy Systems 26 (2018) 937–950.

[8] A.K. Das, S. Das, A. Ghosh, Ensemble feature selection using bi-objective genetic algorithm, Knowledge-Based Systems 123 (2017) 116 – 127.

[9] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155 – 176.

[10] R. Diao, F. Chao, T. Peng, N. Snooke, Q. Shen, Feature selection inspired classifier ensemble reduction, IEEE Transactions on Cybernetics 44 (2014) 1259–1268.

[11] W. Fu, R. Olson, Nathan, G. Jena, PGijsbers, T. Augspurger, J. Romano, P. Saha, S. Shah, S. Raschka, sohnam, DanKoretsky, kadarakos, Jaimecclin, bartdp1, G. Bradway, J. Ortiz, J.J. Smit, J.H. Menke, M. Ficek, A. Varik, A. Chaves, J. Myatt, Ted, A.G. Badaracco, C. Kastner, C. Jerônimo, Hristo, M. Rocklin, R. Carnevale, Epistasislab/tpot: v0.11.5, 2020. URL: `https://doi.org/10.5281/zenodo.3872281`. doi:10.5281/zenodo.3872281.

[12] A. Hadrani, K. Guennoun, R. Saadane, M. Wahbi, Fuzzy rough sets: Survey and proposal of an enhanced knowledge representation model based on automatic noisy sample detection, Cognitive Systems Research 64 (2020) 37 – 56.

[13] Z. Jiang, K. Liu, X. Yang, H. Yu, H. Fujita, Y. Qian, Accelerator for supervised neighborhood based attribute reduction, International Journal of Approximate Reasoning 119 (2020) 122–150.

[14] W. Lee, C.H. Jun, J.S. Lee, Instance categorization by support vector machines to adjust weights in adaboost for imbalanced data classification, Information Sciences 381 (2017) 92 – 103.

[15] H. Lim, D.W. Kim, Pairwise dependence-based unsupervised feature selection, Pattern Recognition 111 (2021) 107663.

[16] J. López, S. Maldonado, M. Carrasco, Double regularization methods for robust feature selection and svm classification via dc programming, Information Sciences 429 (2018) 377 – 389.

[17] M. Mafarja, S. Mirjalili, Whale optimization approaches for wrapper feature selection, Applied Soft Computing 62 (2018) 441 – 453.

[18] A. Mirzaei, Y. Mohsenzadeh, H. Sheikhzadeh, Variational relevant sample-feature machine: A fully bayesian approach for embedded feature selection, Neurocomputing 241 (2017) 181 – 190.

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

[20] X. Qiu, L. Zhang, P. Nagaratnam Suganthan, G.A. Amaratunga, Oblique random forest ensemble via least square estimation for time series forecasting, Information Sciences 420 (2017) 249 – 262.

[21] Y. Qu, C. Shang, N.M. Parthaláin, W. Wu, Q. Shen, Multi-functional nearest-neighbour classification, Soft Computing 22 (2018) 2717–2730.

[22] Y. Qu, C. Shang, Q. Shen, N.M. Parthaláin, W. Wu, Kernel-based fuzzy-rough nearest-neighbour classification for mammographic risk analysis, International Journal of Fuzzy Systems 17 (2015) 471–483.

[23] Y. Qu, Q. Shen, N.M. Parthaláin, C. Shang, W. Wu, Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels, International Journal of Approximate Reasoning 54 (2013) 184 – 195.

[24] Y. Qu, G. Yue, C. Shang, L. Yang, R. Zwiggelaar, Q. Shen, Multi-criterion mammographic risk analysis supported with multi-label fuzzy-rough feature selection, Artificial Intelligence in Medicine 100 (2019) 1–14.

[25] Y. Rubner, C. Tomasi, L.J. Guibas, The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision 40 (2000) 99–121.

[26] C. Shang, Q. Shen, Aiding classification of gene expression data with feature selection: a comparative study, International Journal of Computational Intelligence Research 1 (2005) 68–76.

[27] C.E. Shannon, A mathematical theory of communication, Bell system technical journal 27 (1948) 379–423.

[28] L. Sun, J. Xu, X. Cao, Decision table reduction method based on new conditional entropy for rough set theory, in: 2009 International Workshop on Intelligent Systems and Applications, pp. 1–4.

[29] M. Suo, L. Tao, B. Zhu, X. Miao, Z. Liang, Y. Ding, X. Zhang, T. Zhang, Single-parameter decision-theoretic rough set, Information Sciences 539 (2020) 49 – 80.

[30] C.F. Tsai, Y.C. Chen, The optimal combination of feature selection and data discretization: An empirical study, Information Sciences 505 (2019) 282 – 293.

[31] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: Introduction and review, Journal of Biomedical Informatics 85 (2018) 189–203.

[32] C.Y. Wang, L. Wan, New results on granular variable precision fuzzy rough sets based on fuzzy (co)implications, Fuzzy Sets and Systems (2020).

[33] S. Xia, Z. Zhang, W. Li, G. Wang, E. Giem, Z. Chen, Gbnrs: A novel rough set algorithm for fast adaptive attribute reduction in classification, IEEE Transactions on Knowledge and Data Engineering (2020) 1–1. doi:10.1109/TKDE.2020.2997039.

[34] J. Yang, G. Wang, Q. Zhang, Knowledge distance measure in multigranulation spaces of fuzzy equivalence relations, Information Sciences 448-449 (2018) 18 – 35.

[35] J. Yang, G. Wang, Q. Zhang, H. Wang, Knowledge distance measure for the multigranularity rough approximations of a fuzzy concept, IEEE Transactions on Fuzzy Systems 28 (2020) 706–717.

[36] Y. Yang, M. Loog, A benchmark and comparison of active learning for logistic regression, Pattern Recognition 83 (2018) 401 – 415.

[37] T. Zhang, On the consistency of feature selection using greedy least squares regression, Journal of Machine Learning Research 10 (2009) 555–568.

[38] W. Zheng, F.Y. Wang, C. Gou, Nonparametric different-feature selection using wasserstein distance, in: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 982–988.