Applied Soft Computing Journal 82 (2019) 105543

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

A novel measure of attribute significance with complexity weight

Jinfu Liu *, Mingliang Bai, Na Jiang, Daren Yu

School of Energy Science and Engineering, Harbin Institute of Technology, 150001 Harbin, Heilongjiang, China

HIGHLIGHTS

- A novel measure of attribute significance with complexity weight was defined.
- Structural risk minimization principle was introduced into attribute reduction.
- The number of rules characterizes the complexity of rough set-based classifiers.
- The direct control of complexity in rough set based classification was realized.
- Generalization ability of rough set-based classifiers was significantly improved.

ARTICLE INFO

Article history:
Received 14 March 2018
Received in revised form 30 May 2019
Accepted 30 May 2019
Available online 7 June 2019

Keywords:
Attribute reduction
Attribute significance
Structural risk minimization
Complexity weight
Generalization ability

ABSTRACT

Attribute reduction is one of the most important problems in rough set theory. Conventional attribute reduction algorithms are based on minimal errors in seen objects, namely empirical risk minimization. Classification ability in unseen objects, namely generalization ability is more important in actual applications. Therefore, a good reduct should have good generalization ability. Structural risk minimization (SRM) inductive principle is an effective tool to control the generalization ability of learning machines, which considers complexity and errors in seen objects simultaneously. Therefore, this paper introduces the SRM principle into the definition of attribute significance, proposes that the number of rules can characterize the actual complexity of the rough set-based classifier effectively and defines a novel measure of attribute significance with complexity weight. Based on the new attribute significance, a new heuristic attribute reduction algorithm called HSRM-R algorithm is developed. The 10-fold cross-validation experiments in 21 UCI datasets show that HSRM-R algorithm obtains better generalization ability than conventional attribute reduction algorithms based on dependency degree, information entropy, Fisher score and Laplacian score. Further experiments show that HSRM-R algorithm obtains fewer rules and larger support coefficient. This means HSRM-R algorithm can extract stronger rules, which explains why it has better generalization ability to some extent. Although HSRM-R algorithm consumes more time than conventional algorithms, it obtains optimal classification accuracy in almost all datasets used in the experiments. Thus, the proposed HSRM-R algorithm provides an approach to guaranteeing the generalization ability theoretically in the case where users require high classification accuracy.

1. Introduction

Rough set theory, proposed by Z. Pawlak [1] in 1982, is an effective mathematical tool to handle imprecise, incomplete and uncertain information [2–5]. The main concept of rough set theory is binary equivalence relation or indiscernibility relation on the universe of discourse, which is similar to conflict analysis from the logical point of view. The Pawlak conflict analysis model has proven to be an effective method to handle the problem of conflict in practice [6]. Currently rough set theory has been applied to solve many conflict decision problems [7,8]. Besides, rough set theory has also been successfully applied to many other practical problems including machine learning [9,10], pattern recognition [11,12], fault diagnosis [13,14], decision support systems [15], data mining [16], bilateral filter design [17], linguistic assessment [18], face recognition [19] etc.

Attribute reduction is one of the core problems in rough set theory. Attribute reduction removes the irrelevant, unnecessary or insignificant attributes without violating the abilities of classification for the repository, in order to properly classify the information systems or attain decision-making rules [20]. The definition of attribute significance is important to attribute reduction. Research on attribute significance mainly focuses on three aspects: dependency degree [21], entropy [22–24] and consistency [25,26]. Many heuristic attribute reduction algorithms

* Corresponding author.
E-mail address: jinfuliu@hit.edu.cn (J. Liu).

https://doi.org/10.1016/j.asoc.2019.105543
1568-4946/

have been proposed based on these definitions of attribute significance. To refrain the noise in the actual data, Liu et al. [27] introduced the threshold into attribute reduction and removed the attributes whose significance are lower than a given threshold. Ziarko [28] tolerated the misclassification rate of training data to some degree and proposed variable precision rough sets. Additionally, many researchers focus on minimal attribute reduction. There are two main kinds of algorithms on minimal attribute reduction: discernibility matrix-based algorithms [29–31] and intelligent optimization algorithms including genetic algorithm [32], particle swarm optimization [33], population-based incremental learning algorithm [34] etc. However, Min et al. [35] argued that attribute reduction based on minimal attribute space can obtain smaller rule sets with better predicting performance than minimal attribute reduction. Jia et al. [36] summarized 22 reducts, divided them into six groups, compared the classification accuracy of one typical reduct in each group. They reported that no reducts are optimal in terms of classification accuracy because all 6 compared reducts were not defined based on the classification accuracy measure. Therefore, above definitions of attribute significance and attribute reduction algorithms mainly aim at minimal errors in seen objects, namely empirical risk minimization. Compared with minimal errors in seen objects, minimal errors in unseen objects is more important in the actual applications of rough sets. This is the problem this paper deals with.

Currently, structural risk minimization (SRM) inductive principle proposed by Vapnik and Chervonenkis in [37] is one of the most effective methods to improve the classification ability in unseen objects, namely generalization ability. The SRM principle suggests the generalization ability of learning machines is determined by both errors in seen objects and complexity and that a suitable tradeoff between them can guarantee good generalization ability [38,39]. Based on the SRM principle, support vector machine, a learning machine with good generalization ability, is established and successfully applied in many high-dimensional and nonlinear pattern recognition problems [38–42]. The SRM principle has also been employed to significantly improve the generalization ability of other classifiers such as neural networks [43,44] and decision trees [45,46]. Therefore, this paper introduces the SRM principle into attribute reduction to improve the generalization ability of rough set-based classifier. This paper proposes that the number of rules characterizes the complexity of rough set-based classifier and introduces the complexity weight to adjust the relative proportion of errors in seen objects and complexity. Based on the tradeoff between errors in seen objects and complexity, this paper proposes a novel measure of attribute significance with complexity weight, develops the corresponding attribute reduction algorithm and verifies its superiorities through experiments.

The main contributions of this paper are as follows. Firstly, this paper introduces the SRM principle into attribute reduction to improve the generalization ability of rough set-based classifier. Secondly, this paper discusses the complexity of rough set-based classifier in detail and proposes that the number of rules characterizes the complexity of rough set-based classifier. Thirdly, this paper defines a novel measure of attribute significance based on the SRM principle. The biggest advantage of the new attribute significance is that it considers errors in seen objects and complexity simultaneously while current attribute significance mainly focuses on errors in seen objects.

The rest of this paper is organized as follows. Section 2 briefly reviews the current attribute reduction algorithms and introduces the research orientation. Section 3 briefly introduces the theory for controlling the generalization ability and analyzes the complexity of rough set-based classifier. Section 4 defines a novel measure of attribute significance with complexity weight. Section 5 develops HSRM-R algorithm based on the new attribute significance. Section 6 presents and discusses the computational experiments. Section 7 discusses the superiorities of the proposed algorithm over conventional algorithms. Section 8 concludes the paper.

2. Review of current attribute reduction

2.1. Basic concepts of rough sets

The information system IS is defined as $IS = \langle U, A, V, f\rangle$, where U is the universe, a non-empty set of finite objects, A is the set of attributes, $V = \bigcup_{a\in A} V_a$ is the set of all attribute values and $f: U \times A \to V$ is the information function. For $\forall x \in U$ and $\forall a \in A$, we have $f(x, a) \in V$. The information system that satisfies $A = C \cup D$ and $C \cap D = \varnothing$ is also called the decision table, where C is a set of conditional attributes, and D is a set of decision attributes [32,47].

For any conditional attribute set $B \subseteq A$ in the decision table, the indiscernibility relation $IND(B)$ is defined by $IND(B) = \{(x, y) \in U \times U | f(x, a) = f(y, a), \forall a \in B\}$. For every $x \in U$, the equivalence classes of x, denoted by $[x]_B$, is defined by $[x]_B = \{y \in U | (x, y) \in IND(B)\}$. The family of all equivalence classes of $IND(B)$, i.e., the partition determined by B, is denoted by $U/IND(B)$ or simply U/B. Let X be a subset of U, the B-lower approximation $\underline{B}(X)$ and B-upper approximation $\overline{B}(X)$ of X are defined as follows [47]:

$$\underline{B}(X) = \{x \in U | [x]_B \subseteq X\}, \overline{B}(X) = \{x \in U | [x]_B \cap X \neq \varnothing\}. \quad (1)$$

Let B be a subset of C, the B-positive region $POS_B(D)$ in the relation $IND(B)$ is defined by $POS_B(D) = \bigcup_{X\in U/D} \underline{B}(X)$. The positive region is the sample set which can be undoubtedly classified into a certain class according to the existing attributes. If B is a subset of C such that $POS_B(D) = POS_C(D)$, then B is a D-reduction (reduction with respect to D) of C. The dependency degree or quality of approximation $\gamma_B(D)$ of U/D on B is defined by $\gamma_B(D) = |POS_B(D)| / |U|$, where $|F|$ denotes the cardinality of set F [47].

2.2. Definitions of attribute significance and attribute reduction

Existing definitions of attribute significance are mainly based on three aspects: dependency degree [21], entropy [22–24] and consistency [25,26].

Dependency degree-based definition was first defined by Hu and Cercone [21], which aims to remain the dependency degree of a given decision table unchanged. Let $B \subseteq C$ and $a \in C - B$, then the significance of an individual attribute a added to the set B with respect to the dependency between B and D is represented by $SIG(a, B, D)$, given by:

$$SIG(a, B, D) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D). \quad (2)$$

Entropy based definition was first introduced in 1993/1994 by Skowron [48]. Currently, there are many definitions of entropy, such as Shannon's information entropy [49] Liang's entropy [50,51] and combination entropy [52]. Based on entropy, Jing et al. [22] and Yan et al. [24] defined corresponding attribute significance, which aims to keep the entropy of a given decision table unchanged.

For consistency-based definition, the concept of consistency was proposed by Dash and Liu [25]. Hu et al. [26] defined a measure of attribute significance based on consistency, which aims to the keep the consistency unchanged.

Based on above definitions of attribute significance, heuristic attribute reduction algorithms are proposed to obtain a reduct of

a given decision table $IS = \langle U, A, V, f\rangle$. Yao et al. [53] proposed the generalized deletion method, addition method and addition–deletion method for heuristic attribute reduction algorithms. The deletion method starts with the entire attribute set, checks all attributes for deletion and deletes redundant attribute one by one. The addition method usually starts with an empty set B, and adds the attribute that maximizes the attribute significance SIG into set B in a round until $SIG \leq 0$. The addition–deletion method first uses the addition method to obtain a super-reduce and then deletes the unnecessary attributes in the super-redux until a reduct is found [53].

In actual classification problems, noise is inevitable. Some selected attributes merely fit the noise instead of reflecting the intrinsic characteristics of data. In this case, the classification ability in unseen objects, namely generalization ability cannot be guaranteed. To refrain the noise, Liu et al. [27] modified the heuristic attribute reduction algorithm by introducing a threshold ε to stop the reduction process early, i.e. until $SIG(a, B, D) \leq \varepsilon$ instead of until $SIG(a, B, D) \leq 0$. Details about this algorithm are as follows [27]:

Additionally, Zarko tolerated misclassification in seen objects to some degree and proposed variable precision rough sets [28]. Based on variable precision rough sets, β lower approximate reduct, β upper approximate reduct, β lower distribute reduct and β upper distribute reduct was proposed [54]. Chen et al. [55] presents an incremental algorithm for attribute reduction with variable precision rough sets to address the time complexity of current algorithms.

2.3. Minimal attribute reduction

Many researchers focus on minimal attribute reduction [20, 56]. Now there are mainly two methods of minimal attribute reduction.

One method is based on discernibility matrix [29–31]. Given a decision table $IS = \{U, A = C \cup D, V_a, f\}$, for an object pair (x, y), its discernibility matrix $DM = (DM(x, y))$ is a $|U| \times |U|$ matrix defined by $DM(x, y) = \{a \in C | f(x) \neq f(y) \wedge f_D(x) \neq f_D(y)\}$, where $f_D(x)$ and $f_D(y)$ are the decision of x and y respectively. Based on the discernibility matrix, the discernibility function is defined by $f(DM) = \wedge\{\vee(DM(x, y))|\forall x, y \in U, DM(x, y) \neq \varnothing\}$, where the expression $\wedge\{\vee(DM(x, y))\}$ is the conjunction of all $\vee(DM(x, y))$ and $\vee(DM(x, y))$ is the disjunction of all condition attributes in $DM(x, y)$. Boolean operations are conducted on the discernibility function to search all reducts [29,36]. After finding all reducts, one can select the reduct with fewest attributes from all reducts. The main problem of this method is space and time cost [26]. Another method is based on intelligent optimization algorithm, such as particle swarm optimization [33], genetic algorithm [32] etc. This method transforms the minimal attribute reduction to an optimization problem and aims to obtain a reduct with as few attributes as possible. It reduces the time complexity to some extent but cannot always guarantee the obtained reduct has fewest attributes.

2.4. Minimal attribute space basis for attribute reduction

However, minimal attribute reduction also has some problems. First, finding the minimal reduct is NP-hard [32]. Additionally, for data in reality where attribute domain sizes vary, minimal attribute reduction is unfair since attributes with larger domains tend to have better discernibility or other significance measures and it has severe implications when applied blindly without regarding for the resulting induced concept [57].

Min et al. [35] defined the attribute space as $\prod_{a \in R} |V_a|$, where R is a reduct of the decision table and V_a is the domain of attribute a. They argued that a reduct is optimal if and only if its attribute space is minimal. They reported that minimal attribute space bias for attribute reduction can obtain better generalization performance than minimal attribute reduction.

2.5. Analysis of current attribute reduction algorithm

In the actual applications of the rough set-based classifier, unseen objects need to be classified correctly. Thus the classification ability in unseen objects, namely generalization ability, is more important than the classification ability in seen objects.

Current definitions of attribute significance are based on minimal errors in seen objects without considering the classification accuracy in unseen objects. Additionally, the heuristic attribute reduction algorithm with threshold and variable precision rough sets can refrain noise to some extent by tolerating some errors in seen objects. But they do not consider the classification accuracy in unseen objects systematically. Besides, minimal attribute reduction and minimal attribute space basis reduction aims to obtain good generalization ability by reducing the number of attributes and the attribute space respectively. But the number of attributes and the attribute space are both based on seen objects. Therefore, their generalization ability cannot be guaranteed.

In summary, current reducts are defined based on seen objects, thus the generalization ability cannot be guaranteed.

3. Generalization ability control of rough set-based classifier

3.1. Theory for controlling the generalization ability of learning machines

In the field of machine learning, structural risk minimization (SRM) inductive principle is one of the most effective theories to control the generalization ability [38,39]. This part will brief introduce relevant theories about the SRM principle.

The essence of learning problem is to minimize the expected risk function given by $R(\alpha) = \int Q(z, \alpha) dP(z)$, where $P(z)$ is the probability function and $Q(z, \alpha)$ is the loss function between predicted values and actual values. In reality, $P(z)$ is unknown but an independent identically distributed sample $z_1, z_2, \ldots, z_l$ is given. Thus, the expected risk function is usually replaced by the following empirical risk function $R_{emp}(\alpha)$ based on errors in seen objects [38].

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha) \tag{3}$$

where l is the sample size. The empirical risk represents errors in seen objects.

Errors in unseen objects is directly determined by the expected risk. Small expected risk can guarantee small classification errors in seen objects, namely good generalization ability. There exists the following relationship between the expected risk and the empirical risk. Let Λ be a set of totally bounded functions. If $Q(z, \alpha)$ satisfies $0 \leq Q(z, \alpha) \leq B, \alpha \in \Lambda$, then the following inequality holds with probability at least $1 - \eta$ simultaneously for all loss functions $Q(z, \alpha) \leq B, \alpha \in \Lambda$ (including the function that minimizes the empirical risk) [39]:

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\xi}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha)}{B\xi}}\right), \tag{4}$$

where the first summand on the right side is the empirical risk, the second summand is the confidence interval that depends on the complexity. In the case where the set of functions contains

Algorithm: HRS Algorithm
Input: Decision table $IS = <U, A = C \bigcup D, V, f>$ and a threshold $\varepsilon \in [0,1]$
Output: One reduct B
Begin
Compute the quality of approximation $\gamma_C(D)$;
$B \leftarrow \varnothing$; // B is a pool to contain the selected attributes.
while $B \subset C$ **do**
  for each $a \in C - B$ **do**
    Compute $SIG(a, B, D)$;
    Select a_{max} such that $SIG(a_{max}, B, D)$ is maximum;
  end
  $B \leftarrow B \bigcup \{a_{max}\}$;
  if $SIG(a_{max}, B, D) \le \varepsilon$ **then** exit the loop;
end
for each $a \in B$ **do**
  if $\gamma_B(D) - \gamma_{B-\{a\}}(D) \le \varepsilon$ **then** $B \leftarrow B - \{a\}$;
end
return B;
end

an infinite number of elements and possess a finite Vapnik–Cervonenkis (VC) dimension h, h is the complexity of the approximating function and the expression for ξ in inequality (4) is given in Eq. (5). In the case where the set of loss functions $Q(z, \alpha)$ contains a finite number N of elements, N is the complexity of the approximating function and the expression for ξ is given in Eq. (6) [38].

$$\xi = 4\frac{h(\ln\frac{2l}{h} + 1) - \ln(\frac{\eta}{4})}{l} \tag{5}$$

$$\xi = 2\frac{\ln N - \ln\eta}{l} \tag{6}$$

According to Inequality (4), a small value of empirical risk provides a small value of expected risk when the sample size l is large. However, in the case of small samples sizes, even a small $R_{emp}(\alpha)$ does not guarantee small expected risk. Therefore, the empirical risk is merely an approximation of the expected risk in the case of large sample sizes. Only by minimizing the expected risk can guarantee the generalization ability. To minimize the expected risk, both of the two terms on the right side of Inequality (4) should be minimized. However, the two terms are contradictory. As the complexity of the approximating function increases, the confidence interval increases monotonously while the empirical risk decreases monotonously. Thus, the SRM principle suggests a tradeoff between the empirical risk and complexity to minimize the expected risk. Controlling the complexity on the precondition of small empirical risk can guarantee small expected risk and good generalization ability. Therefore, the complexity must be controlled while decreasing the empirical risk to improve the generalization ability.

3.2. Complexity metric of rough set-based classifier

In the rough set-based classifier, the number of possible values of conditional attributes and decisional attributes is finite, so the set of loss functions has a finite number N of elements. Thus, the complexity of rough set-based classifier is measured by N.

Without loss of generalization, it is assumed that there is a decision table with n conditional attributes, one decisional attribute and N_D possible decision values. According to counting principle, the values of all conditional attributes have $\prod_{i=1}^{n}\left|V_{a_i}\right|$ possible combinations in the rough set-based classifier, where $\left|V_{a_i}\right|$ is the number of possible values of the conditional attribute a_i. The expression $\prod_{i=1}^{n}\left|V_{a_i}\right|$ is called attribute space in [35]. Let N_p equal $\prod_{i=1}^{n}\left|V_{a_i}\right|$. Then these combinations divide the decision space into N_p partitions. Each partition is described by $P(a_1, a_2, \ldots, a_n) = (a_1 = v_{a_1}) \wedge (a_2 = v_{a_2}) \ldots \wedge (a_n = v_{a_n})$, where v_{a_i} is the element in set V_{a_i}. In the case where all the conditional attributes have the same number of possible values, it holds that $N_p = |V_a|^n$ and the number of conditional attributes can represent the attribute space to some extent.

However, the attribute space only reflects the maximal possible number of partitions of decision space. The actual number of partitions may be much lower when the classification process is considered. For example, if it holds that $n = 2$, $\left|V_{a_1}\right| = 2$ and $\left|V_{a_2}\right| = 3$ in the above decision table, then the decision space is divided into $2 \times 3 = 6$ partitions shown in Fig. 1(a). If the decision is "class 1" for all seen objects when "$a_1 = 1$", then a_2 is redundant in terms of classification. Thus, as is illustrated in Fig. 1(b), the three partitions can be merged into one partition. Thus, N_P merely reflect the maximal number of partitions rather than the actual number of partitions. The actual number of partitions of decision space is much lower than N_P.

If we assign a decision value to the four partitions in Fig. 1(b) respectively, then each partition with a decision value forms a rule. The number of actual partitions of decision space equals the number of rules, denoted by N_R. All rules constitute a set of rules. A set of rules can make decisions for objects, so it is also called a decision function. All decision functions constitute the set of decision functions. The number of decision functions is equal to the number of ways to assign decision values. In a classification problem with N_D different decision values, there are $N_D^{N_R}$ possible assignment ways, so there are $N_D^{N_R}$ decision functions. Usually, each decision function corresponds to one loss function. Therefore, the actual number N of elements that the set of loss function actually contains is $N_D^{N_R}$ in rough set-based classifier. In a given the rough set-based classifier, N_D is a constant. Thus, the number of rules N_R characterizes the actual complexity of a given rough set-based classifier. In summary, the

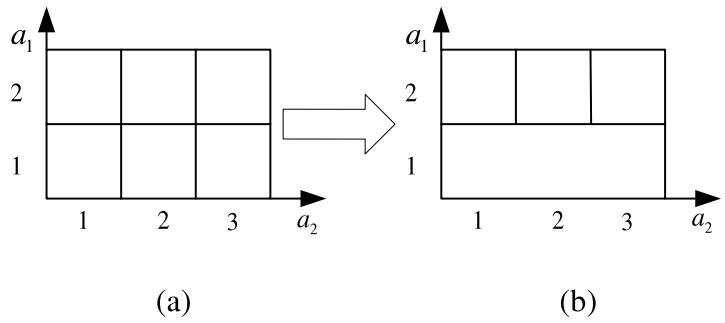


Fig. 1. The division of decision space.

number of rules characterizes the actual complexity of rough set-based classifier while the attribute space merely characterizes the maximal possible complexity.

4. New attribute significance with complexity weight

As is illustrated in Section 3, the SRM principle suggests that the empirical risk and complexity should be considered simultaneously to guarantee good generalization ability. According to inequality (5), $R_{emp} + B\xi(1 + \sqrt{1 + 4R_{emp}(\alpha)/B\xi})/2$ is equal to or greater than the expected risk with probability at least $1 - \eta$. Thus, minimizing $R_{emp} + B\xi(1 + \sqrt{1 + 4R_{emp}(\alpha)/B\xi})/2$ can minimize the expected risk. The first summand R_{emp} is the empirical risk, which can be described by misclassification rate. The second summand $B\xi(1 + \sqrt{1 + 4R_{emp}(\alpha)/B\xi})/2$ depends on complexity, which is a monotonous increasing function of ξ. The variable ξ is a monotonous increasing function of $\ln N/l$, so $B\xi(1 + \sqrt{1 + 4R_{emp}(f)/B\xi})/2$ is also a monotonous increasing function of $\ln N/l$. Thus, it can be controlled by $\ln N/l$. To adjust the relative proportion of empirical risk and relative complexity, a controllable weight coefficient w is introduced. Therefore, the structural risk R_{stru} is defined by the empirical risk plus weighted complexity, namely the following equation. Minimizing R_{stru} can guarantee the generalization ability.

$$R_{stru} = R_{emp} + w\frac{\ln N}{l} \tag{7}$$

In a decision table $IS = \langle U, A = C \cup D, V, f\rangle$, let B be a subset of conditional attribute set C. The dependency degree $\gamma_B(D)$ is the ratio of objects that can be correctly classified. Large $\gamma_B(D)$ means small empirical risk. Thus, the empirical risk of the rough set-based classifier can be characterized by $1 - \gamma_B(D)$. It has been proposed in Section 3.2 that the number N in the rough set-based classifier equals to $N_D^{N_R}$. Thus, $\ln N/l$ equals to $N_R \ln N_D/l$. In a given rough set-based classifier, the number of classes N_D is a constant. Therefore, the constant term $\ln N_D$ is merged into the weight coefficient w and the complexity of rough set-based classifier is controlled by wN_R/l. If the conditional attribute subset B is used for classification and the number of rules is $N_R(B)$, then the complexity is controlled by $wN_R(B)/l$. Thus, the structural risk $R_{stru}(B)$ of rough set-based classifier is characterized by the following equation.

$$R_{stru}(B) = 1 - \gamma_B(D) + w\frac{N_R(B)}{l} \tag{8}$$

For a conditional attribute subset used for classification, small structural risk means good generalization ability. Thus, a conditional attribute subset with small structural risk should have large significance to guarantee generalization ability. The structural risk of a conditional attribute subset B is small if it has a large value of $\gamma_B(D)$ and a small value of $wN_R(B)/l$. Thus, the significance of conditional attribute subset B, denoted by $SIG_{stru}(B)$, is defined as follows:

$$SIG_{stru}(B) = \gamma_B(D) - w\frac{N_R(B)}{l}, \tag{9}$$

where $\gamma_B(D)$ represents the contribution of subset B to classification and $wN_R(B)/l$ represents its relative complexity.

The practical meaning of $SIG_{stru}(B)$ is as follows. The number of extracted rules is always is less than or equal to the sample size l. Meanwhile, it holds that $\gamma_B(D) \in [0, 1]$. Thus, it holds that $SIG_{stru}(B) \in [-w, 1)$. When $\gamma_B(D) = 1$ and $N_R(B)/l$ is very small, $SIG_{stru}(B)$ is very close to 1. In this case, all seen objects can be classified correctly, the extracted rules are strong ones that reflect the intrinsic characteristics of data well and the generalization ability are good. When $\gamma_B(D) = 0$ and $N_R(B)/l = 1$, $SIG_{stru}(B)$ equals $-w$. In this case, no seen objects can be classified correctly, the extracted rules are weak ones and the generalization ability are poor. Thus, the conditional attribute subset B with large $SIG_{stru}(B)$ usually has large dependency degree and few rules, and thus has small structural risk and good generalization ability.

In the process of attribute reduction, the conditional attribute subset becomes $B \cup \{a\}$ when a new attribute a that satisfies $a \in C - B$ is added to the set B. If conditional attribute subset $B \cup \{a\}$ is used for classification and the number of rules is $N_R(B \cup \{a\})$, then the significance of conditional attribute subset $B \cup \{a\}$ is characterized by the following equation.

$$SIG_{stru}(B \cup \{a\}) = \gamma_{B\cup\{a\}}(D) - w\frac{N_R(B \cup \{a\})}{l} \tag{10}$$

In the process of attribute reduction, when a new attribute a is added to the set B, the significance of attribute a, denoted by $SIG_{stru}(a, B, D)$, is defined by the significance of set $B \cup \{a\}$ minus the significance of set B. Therefore, the novel measure of significance is defined as follows:

$$\begin{aligned} SIG_{stru}(a, B, D) &= SIG_{stru}(B \cup \{a\}) - SIG_{stru}(B) \\ &= [\gamma_{B\cup\{a\}}(D) - \gamma_B(D)] - w[\tfrac{N_R(B\cup\{a\})}{l} - \tfrac{N_R(B)}{l}], \end{aligned} \tag{11}$$

where $\gamma_{B\cup\{a\}}(D) - \gamma_B(D)$ represents the contribution of attribute a to classification and $w[N_R(B \cup \{a\})/l - N_R(B)/l]$ represents the increase of the relative complexity caused by the attribute a. The new measure of attribute significance considers both empirical risk and complexity.

5. Attribute reduction based on new attribute significance

Based on the proposed attribute significance, this paper proposes a new heuristic attribute reduction algorithm.

To compute the proposed attribute significance, the number of rules N_R needs to be computed first. Among many known rule extraction algorithms, LEM2 algorithm proposed by Grzymala-Busse in [58] is one of the most common algorithms in the rough set-based classifier. Thus, this paper used LEM2 algorithm to extract rules. Details about LEM2 algorithm can be seen in [58].

Table 1
Main characteristics of the datasets used in the experiments.

ID	Datasets	Instances	Conditional attributes	Classes
1	Hepatitis (Hep.)	155	19	2
2	Iono	351	34	2
3	Horse	368	22	2
4	Votes	435	16	2
5	Credit	690	15	2
6	Zoo	101	16	7
7	Lymphography (Lym.)	148	18	4
8	Wine	178	13	3
9	Flags	194	28	8
10	Autos	205	23	6
11	Images	210	19	7
12	Soybean	683	35	19
13	Vehicle	846	18	4
14	Tic	958	9	2
15	German	1000	24	2
16	Anneal	898	38	3
17	Semi-Bumps (Bumps)	2584	18	2
18	Burst Header Packet (BHP)	1075	21	4
19	Diabetic Retinopathy Debrecen (DRD)	1151	19	2
20	Mushroom	8124	22	7
21	Page Block (PB)	5473	10	5

In heuristic attribute reduction algorithms, there are three search strategies including deletion method, addition method and addition–deletion method. Here, the addition method is used, and the other two methods can also be used with a slight modification of the addition method. Let B start with an empty set, the attribute a_k that satisfy $a_k \in C - B$ with maximal attribute significance $SIG_{stru}(a_k, B, D)$ is selected each time. When the contribution of attribute a_k to classification is offset by the increase of complexity caused by it, $SIG_{stru}(a_k, B, D)$ is zero. The minimal expected risk is obtained, and the algorithm stops at this time. Formally, the new heuristic attribute reduction algorithm called HSRM-R algorithm is as follows.

The main difference between HSRM-R algorithm and conventional HRS algorithm lies in the measure of attribute significance. HSRM-R algorithm and HRS algorithm use $SIG_{stru}(a, B, D)$ and $SIG(a, B, D)$ as attribute significance respectively. The proposed attribute significance $SIG_{stru}(a, B, D)$ considers both complexity and errors in seen objects while $SIG(a, B, D)$ only considers errors in seen objects.

A reduct of the original decision table B and a rule set $\mathbb{T}$ can be obtained by HSRM-R Algorithm. To evaluate discovered rules and predict the unseen objects, the support coefficient $\mu_{sup}(r)$ is defined. For a decision table $IS = \langle U, A = C \cup D, V, f\rangle$, if $Q \subseteq A$ and $E \in U/Q$, then $Des(E, Q) = \wedge(a, f(x, a))$, where $x \in E$, $a \in Q$, f is the information function $f: U \times A \to V$ and $(a, f(x, a))$ denotes that the value of the attribute a equals $f(x, a)$, is called the description of class E with respect to Q. Note that for $\forall x \in U$ and $\forall a \in A$, we have $f(x, a) \in V$. Next, the definition of $Des(X, B)$ and $Des(Y, D)$ are given. Firstly, let $B \subseteq C$ and $X \in U/B$, then $Des(X, B) = \wedge(a, f(x, a))$, where $x \in X$, $a \in B$, f is the information function $f: U \times A \to V$ and $(a, f(x, a))$ denotes that the value of the conditional attribute a equals $f(x, a)$, is called the description of class X with respect to B. Secondly, let $Y \in U/D$, then $Des(Y, D) = \wedge(d, f(y, d))$, where $y \in Y$, $d \in D$, f is the information function $f: U \times A \to V$ and $(d, f(y, d))$ denotes that the value of the decisional attribute d equals $f(y, d)$, is called the description of class Y with respect to D. Then the support coefficient of a decision rule $r: Des(X, B) \to Des(Y, D)$, can be defined as follows [27]:

$$\mu_{sup}(r) = \frac{|X \cap Y|}{|U|} \tag{12}$$

The support coefficient $\mu_{sup}(r)$ can measure the number of objects that rules can cover. Thus, it measures the strength of the rules. A large support coefficient means strong rules. Therefore, the support coefficient is used to make decisions for unseen objects. If there are n rules $r_1, r_2, \ldots, r_n$ matching with the description of the unseen objects and there are m different decisions $d_1, d_2, \ldots, d_m$, then the overall support coefficient of rules with decision d_j is computed as follows :

$$\mu_{overall}(d_j) = \sum_{r_i \to d_j} \mu_{sup}(r_i) \tag{13}$$

where $r_i \to d_j$ means that the decision of rule r_i is d_j and $\mu_{sup}(r_i)$ is the support coefficient of rule r_i. According to the principle of majority voting, the decision algorithm can give the decision of an unseen object with the following guideline [27]:

$$d: \mu_{overall}(d) = \max_j \mu_{overall}(d_j) \tag{14}$$

6. Experimental evaluations

To verify the superiorities of the proposed HSRM-R algorithm, this paper performed 10-fold cross-validation experiments in 21 UCI datasets [59] outlined in Table 1. All experiments were coded in MATLAB and executed in a PC Intel Core i5-8400 with 2.80 GHz and 8 GB of RAM running Windows 10 (64 bits).

6.1. Comparison between HSRM-R algorithm and conventional HRS algorithm

6.1.1. Selection of optimal complexity weight in HSRM-R algorithm

In the proposed HSRM-R algorithm, the complexity weight w is an important parameter, representing the relative proportion of empirical risk and complexity. To select the optimal w, a set $\{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ was given and 10-fold cross-validation experiment was performed. Experimental results are shown in Table 2. In Table 2, the bold values represent the optimal classification accuracy. As w increases, the classification accuracy in most datasets first increases then decreases. When w is 0.5, the optimal classification accuracy is obtained in 12 datasets, the classification accuracy is close to the optimal value in the other datasets and the mean classification accuracy is also optimal. When w is 1, the optimal classification accuracy is obtained in 10 datasets and the classification accuracy is close to the optimal value in the other datasets. Therefore, the optimal complexity weight w is 0.5 and the suboptimal complexity weight w is 1.

Algorithm: HSRM-R algorithm

Input: $IS = < U, A = C \cup \boldsymbol{D}, \boldsymbol{V}, \boldsymbol{f} >$ and a weight coefficient $w \in [0, +\infty)$

Output: One reduct B and one rule set $\mathbb{T}$

Step 1: $B \leftarrow \varnothing$; // B is the pool to contain the selected attributes.

$\mathbb{T} \leftarrow \varnothing$; // $\mathbb{T}$ is the pool to contain the extracted rules.

Step 2: **for** each $a_i \in C - B$

Extract rules from set B and obtain rule set $\mathbb{T}$ as well as the number of rules;

Extract rules from set $B \cup \{a_i\}$ and obtain the number of rules $N_R(B \cup \{a_i\})$;

Compute $SIG_{stru}(a_i, B, D)$ using the following equation:

$$SIG_{stru}(a_i, B, D) = [\gamma_{B \cup \{a_i\}}(D) - \gamma_B(D)] - w[\frac{N_R(B \cup \{a_i\})}{l} - \frac{N_R(B)}{l}];$$

end

Step 3: Select the attribute a_k which satisfies:

$SIG_{stru}(a_k, B, D) = \max_i(SIG_{stru}(a_i, B, D))$;

Step 4: **if** $SIG_{stru}(a_k, B, D) \geq 0$

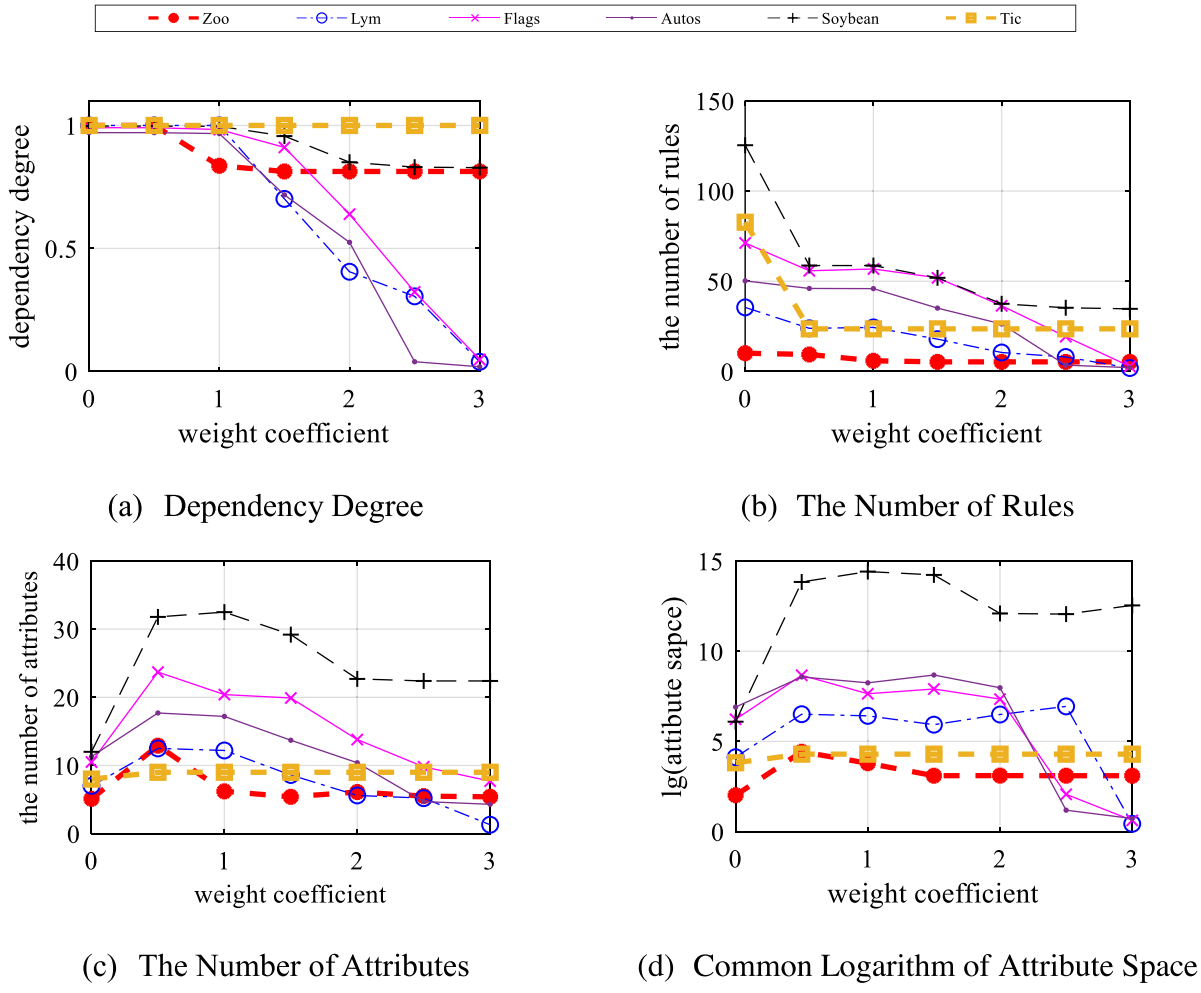
$B \leftarrow B \cup \{a_k\}$;

Go to Step 2;

else

Return B and $\mathbb{T}$;

Step 5: **end**



(a) Dependency Degree (b) The Number of Rules

(c) The Number of Attributes (d) Common Logarithm of Attribute Space

Fig. 2. The change of different metrics with complexity weight.

Table 2
Classification accuracy of HSRM-R algorithm under different w.

Datasets	0	0.5	1	1.5	2	2.5	3
Hep.	0.8325	0.8775	**0.8971**	0.8663	0.8529	0.8333	0.8583
Iono	0.9058	0.9201	0.9172	0.9258	0.9145	**0.9314**	0.9286
Horse	0.9616	**0.9700**	**0.9700**	0.9646	0.9592	0.9673	0.9673
Votes	0.9451	**0.9679**	0.9633	0.9587	0.9540	0.9586	0.9564
Credit	**0.8261**	0.8246	**0.8261**	0.8203	0.8217	0.8203	**0.8261**
Zoo	**0.9418**	**0.9418**	0.8618	0.8909	0.8909	0.8909	0.8909
Lym.	0.7848	**0.8043**	**0.8043**	0.7581	0.6290	0.6157	0.5676
Wine	0.9382	**0.9549**	**0.9549**	0.9438	0.9382	0.9271	0.9382
Flags	0.6142	**0.6389**	0.6189	0.6134	0.5671	0.4645	0.3411
Autos	0.7333	**0.7645**	**0.7645**	0.6664	0.5619	0.3474	0.3574
Images	0.8667	**0.8714**	0.8524	0.8667	0.8667	0.8619	0.8619
Soybean	0.8198	**0.9253**	0.9238	0.9062	0.8856	0.8812	0.8855
Vehicle	0.6666	0.6748	**0.6819**	0.6784	0.6759	0.5913	0.6008
Tic	0.7777	**0.9896**	**0.9896**	**0.9896**	**0.9896**	**0.9896**	**0.9896**
German	0.6950	0.6990	0.6990	**0.7010**	0.6970	0.7000	0.7000
Anneal	0.8809	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Bumps	0.9214	0.9133	0.9129	0.9141	0.9141	0.9160	**0.9222**
BHP	0.9925	0.9925	0.9925	0.9925	**0.9934**	**0.9934**	0.9925
DRD	0.6125	0.6125	**0.6134**	0.6021	0.5917	0.5847	0.5700
Mushroom	0.4846	**0.6290**	0.6273	0.6262	0.6262	0.6262	0.6262
PB	0.9614	0.9614	0.9614	0.9614	0.9614	0.9618	**0.9625**
Mean	0.8173	**0.8540**	0.8492	0.8403	0.8234	0.8030	0.7973

Table 3
Classification accuracy obtained by conventional HRS algorithm under different ε.

Datasets	0	0.05	0.10	0.15	0.20	0.25
Hep.	0.8325	0.8650	0.8721	0.8721	0.8913	**0.9100**
Iono	**0.9058**	0.8888	0.8944	0.8973	0.8973	0.8944
Horse	0.9616	**0.9700**	0.9646	0.9673	0.9673	0.9673
Votes	0.9451	0.9609	**0.9725**	0.9656	0.9656	0.9633
Credit	**0.8261**	0.8232	0.8145	0.8101	0.8014	0.8130
Zoo	0.9418	**0.9509**	0.9409	0.8909	0.8909	0.8609
Lym.	0.7848	0.7852	0.7852	0.7790	**0.7857**	**0.7857**
Wine	0.9382	0.9333	**0.9386**	**0.9386**	**0.9386**	**0.9386**
Flags	**0.6142**	0.5787	0.5937	0.6095	0.5987	0.5826
Autos	0.7333	**0.7686**	0.7488	0.7495	0.7355	0.7210
Images	**0.8667**	0.8333	0.8286	0.8476	0.8476	0.8381
Soybean	0.8198	**0.8945**	0.8930	0.8915	0.8856	0.8842
Vehicle	0.6772	**0.6820**	0.6689	0.6677	0.6702	0.6666
Tic	**0.9373**	0.8956	0.8956	0.8956	0.8403	0.7777
German	**0.6990**	**0.6990**	**0.6990**	**0.6990**	**0.6990**	0.6950
Anneal	**1.0000**	**1.0000**	**1.0000**	0.9555	0.9555	0.8809
Bumps	0.9137	0.9152	0.9176	0.9180	**0.9226**	0.9214
BHP	**0.9925**	0.9680	0.9502	0.9341	0.9031	0.8889
DRD	0.6125	**0.6169**	0.6012	0.5804	0.5856	0.5873
Mushroom	0.4846	0.5793	0.5793	0.5820	**0.6115**	0.5990
PB	**0.9614**	**0.9614**	0.9552	0.9560	0.9538	0.9536
Mean	0.8309	**0.8367**	0.8340	0.8289	0.8261	0.8157

To explore why 0.5 is the optimal complexity weight in our experiment, the change of different metrics with w is studied in the paper. Results are shown in Fig. 2(a)–(d).

Shown in Fig. 2(a), when w increases from 0 to 0.5, the dependency degree almost remains unchanged, which means the empirical risk almost remains unchanged. When w increases from 0.5 to 1, the dependency degree in zoo dataset decreases rapidly but almost remains unchanged in other datasets. When w exceeds 1, the dependency degree in many datasets begins to decrease sharply, which means the empirical risk increases sharply. Therefore, small empirical risk can be obtained when w is less than 1. Meanwhile, in Fig. 2(b), the number of rules, namely complexity, decreases rapidly when w increases from 0 to 0.5. But it decreases slowly or remains almost unchanged with the further increase of w. The best generalization ability can be obtained by controlling the complexity on the precondition of relatively small empirical risk. When w is between 0.5 and 1, the empirical risk and complexity are both small. The empirical risk increases sharply while the complexity decreases slowly with the further increase of w, which leads to bad generalization ability. Therefore, the optimal w should be between 0.5 and 1. Additionally, the empirical risk begins to increase when w is 0.5 in some datasets. Thus, the classification accuracy when w is 0.5 is usually better than the classification accuracy when w is 1. This explains why the optimal and suboptimal classification accuracy are obtained when w is 0.5 and 1 respectively.

Additionally, this paper compared the change trend of attribute space, the number of attributes and the number of rules to verify that the number of rules characterizes the actual complexity. Shown in Fig. 2(b)–(d), when w increases from 0 to 1, the number of attributes and the attribute space increase while the number of rules decreases. Most attributes and largest attribute space are obtained when w is between 0.5 and 1. Interestingly, the optimal classification accuracy is also obtained at this time. The reason is that the proposed HSRM-R algorithm may need more attributes and larger attribute space to search rules with better characterizing ability. The number of attributes and the attribute space begin to decrease, and the number of rules also decreases with the further increase of w. Thus, the change trend of the attribute space is similar to the number of attributes but different from the number of rules. The inconsistency indicates attribute space and the number of attributes merely reflects the maximal possible complexity instead of actual complexity to some extent.

6.1.2. Selection of optimal threshold in convention HRS algorithm

In HRS algorithm, a set $\{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$ is given and 10-fold cross-validation was performed to select the optimal threshold ε. Table 3 shows the classification accuracy of HRS algorithm under different thresholds. When the threshold is 0.05, the optimal classification accuracy can be obtained in 9 datasets, and the mean classification accuracy is also optimal. Although the optimal classification accuracy in 9 datasets can also be obtained under the threshold of 0, the classification accuracy in other datasets are mostly worse than that under the threshold of 0.05. Therefore, the optimal ε in HRS algorithm is 0.05.

6.1.3. Comparison between HSRM-R algorithm and HRS algorithm

After above parameter selection, this paper used the optimal parameters of the two algorithms, namely $\varepsilon = 0.05$ and $w = 0.5$, for attribute reduction and performed 10-fold cross-validation experiment again to compare the performance of the two algorithms. The classification accuracy, the dependency degree, the number of rules, the number of attributes, attribute space, support coefficient and rule length of HSRM-R algorithm and HRS algorithm are compared in Table 4.

In Table 4, although HSRM-R algorithm obtains more attributes and larger attribute space than conventional HRS algorithm, it obtains fewer rules. The number of rules rather than the number of attributes or attribute space characterizes the actual complexity directly. Thus, HSRM-R algorithm obtains smaller complexity. Additionally, HSRM-R algorithm obtains larger support coefficient, which means it can extract stronger rules. Meanwhile, HSRM-R algorithm has a larger dependency degree than the conventional HRS algorithm, which means smaller empirical risk. Smaller empirical risk and smaller complexity mean smaller structural risk. Therefore, HSRM-R algorithm obtains better classification accuracy and better generalization ability than the conventional HRS algorithm.

6.1.4. Comparison between HRS and HSRM-R under similar dependency degree

In Table 4, the dependency degree of HRS algorithm is obviously smaller than HSRM-R algorithm. Thus, its empirical risk is larger than HSRM-R algorithm. To further verify the effectiveness of introducing complexity control, the two algorithms should be compared under the same empirical risk. Thus, the threshold

Table 4
Comparison between HRS Algorithm and HSRM-R algorithm.

Datasets	Accuracy		Dependency degree		Rule number		Support coefficient	
	HSRM-R	HRS	HSRM-R	HRS	HSRM-R	HRS	HSRM-R	HRS
Hep.	**0.9163**	0.8713	**1.0000**	0.9599	**13.2**	16.8	**0.1544**	0.1170
Iono	**0.9288**	0.8917	**1.0000**	0.9715	**27.1**	38.0	**0.0744**	0.0434
Horse	**0.9728**	0.9649	**1.0000**	0.9940	**13.1**	22.2	**0.1445**	0.0715
Votes	**0.9679**	0.9566	**0.9946**	0.9625	**17.1**	17.5	**0.1368**	0.1178
Credit	**0.8333**	0.8087	**0.9643**	0.9233	**107.4**	126.7	**0.0212**	0.0161
Zoo	**0.9500**	**0.9500**	**1.0000**	0.9956	**9.2**	10.3	**0.1119**	0.0993
Lym.	**0.8243**	0.8043	**1.0000**	0.9655	**25.1**	39.0	**0.0631**	0.0385
Wine	**0.9608**	0.9382	**1.0000**	0.9663	**9.3**	12.5	**0.1610**	0.1005
Flags	**0.6500**	0.5832	**0.9897**	0.9565	**56.0**	73.5	**0.0243**	0.0174
Autos	**0.7907**	0.7171	**0.9664**	0.9382	**48.6**	54.8	**0.0259**	0.0222
Images	**0.8905**	0.8476	**0.9783**	0.9481	**22.8**	27.4	**0.0501**	0.0405
Soybean	**0.9253**	0.8828	**0.9974**	0.9645	**59.2**	86.9	**0.0205**	0.0141
Vehicle	**0.6714**	0.6655	**0.8551**	0.8241	**172.4**	185.7	**0.0091**	0.0083
Tic	**0.9917**	0.8966	**1.0000**	0.9921	**23.7**	120.2	**0.0445**	0.0125
German	**0.7030**	**0.7030**	**0.0249**	**0.0249**	**13.2**	**13.2**	**0.1944**	**0.1944**
Anneal	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**5.0**	**5.0**	**0.2000**	**0.2000**
Bumps	0.9133	**0.9145**	**0.4643**	0.4517	**148.9**	149.1	**0.0090**	**0.0090**
BHP	**0.9962**	0.9755	**0.9962**	0.9694	**46.8**	64.7	**0.0286**	0.0181
DRD	**0.6377**	0.6351	**0.2161**	0.2099	72	**69.7**	0.0273	**0.0274**
Mushroom	**0.6273**	0.5865	0.4852	**0.5445**	**214.2**	776	**0.0196**	0.0015
PB	**0.9616**	0.9585	**0.9648**	0.9298	**201.0**	204.7	**0.0152**	0.0105
Mean	**0.8625**	0.8358	**0.8523**	0.8330	**62.2**	100.7	**0.0731**	0.0562

Datasets	Attribute		Common logarithm of attribute space		Rule length	
	HSRM-R	HRS	HSRM-R	HRS	HSRM-R	HRS
Hep.	15.4	8.3	4.5	2.8	3.30	3.26
Iono	15.1	6.2	12.4	4.2	2.70	2.69
Horse	9.5	3.0	5.0	1.9	2.18	2.17
Votes	11.6	7.0	4.0	2.1	3.47	3.45
Credit	14.7	9.1	6.5	4.5	4.40	4.00
Zoo	13.6	4.9	4.8	2.0	2.21	2.25
Lym.	11.7	5.1	6.2	3.3	2.79	2.59
Wine	8.3	3.1	4.2	1.8	2.10	2.11
Flags	22.9	8.0	8.6	5.4	3.25	2.77
Autos	17.7	8.7	8.7	5.7	2.86	2.82
Images	17.2	5.3	8.3	3.7	2.55	2.44
Soybean	30.6	10.7	13.8	5.6	3.98	4.03
Vehicle	17.6	12.4	10.2	7.1	4.47	4.42
Tic	9.0	7.0	4.3	3.3	3.66	4.44
German	24.0	5.1	1.6	1.6	3.02	3.02
Anneal	35.2	3.0	8.0	1.1	1.24	1.56
Bumps	17.9	8.8	6.0	4.7	4.10	4.07
BHP	17.4	4	19.3	5.8	2.03	1.70
DRD	19.0	12.3	5.1	4.6	4.19	4.17
Mushroom	18.1	11	11.7	8.1	4.54	7.30
PB	9.9	5.3	8.0	4.5	3.37	3.24
Mean	17.0	7.1	7.7	4.0	3.16	3.26

ε in HRS algorithm is set to be 0 to ensure its dependency degree is equal to or greater than HSRM-R algorithm. Comparison between them are shown in Table 5. Shown in Table 5, although HRS algorithm with adjusted ε has larger dependency degree than HSRM-R algorithm, its classification accuracy is worse than HSRM-R algorithm in almost all datasets. This further proves that the complexity control is an effective way to control the generalization ability. Thus, the proposed algorithm can guarantee the generalization ability effectively.

6.2. Comparison between HSRM-R algorithm and other conventional algorithms

Currently there are many famous and classical attribute reduction algorithms besides HRS algorithm, such as information entropy based algorithm [60], Fisher Score [61,62], Laplacian Score [63] etc. The information entropy based algorithm uses information entropy to evaluate the attribute significance. It usually uses heuristic attribute reduction algorithm like HRS algorithm. This paper calls it HE algorithm. Details about HE algorithm can be seen in literature [60]. Both Fisher score (FS) and Laplacian score (LS) require users to compute a score for each attribute independently and select the top-m ranked features with large scores. The difference between FS and LS lies in the definition of attribute scores. Details about FS algorithm and LS algorithm can be seen in literature [61] and literature [63] respectively. All the three algorithms are commonly-used in attribute reduction. Therefore, this paper compared the proposed HSRM-R algorithm with these algorithms besides HRS algorithm. The parameters of the three algorithms in the experiment are set as follows. For HE algorithm, this paper uses the same ε as HRS algorithm, namely $\varepsilon = 0.05$. For FS and LS algorithm, this paper uses the same parameter as literature [61], namely the number of selected features is 50% of the dimensionality of the data. After attribute reduction, LEM2 algorithm is used for classification. The classification accuracy of 10-fold cross-validation is shown in Table 4.

Shown in Table 4, HSRM_R algorithm obtains the optimal classification accuracy in 19 datasets and its mean classification accuracy is also optimal. Therefore, HSRM-R algorithm has the best classification accuracy and generalization ability among the five algorithms.

Table 5
Comparison between HRS and HSRM-R under similar dependency degree.

Datasets	Dependency degree		Classification accuracy	
	HRS	HSRM-R	HRS	HSRM-R
Hep.	**1.0000**	**1.0000**	0.8775	**0.9163**
Iono	**1.0000**	**1.0000**	0.9174	**0.9288**
Horse	**1.0000**	**1.0000**	0.9564	**0.9728**
Votes	**0.9959**	0.9946	0.9542	**0.9679**
Credit	**0.9646**	0.9643	0.8217	**0.8333**
Zoo	**1.0000**	**1.0000**	**0.9600**	0.9500
Lym.	**1.0000**	**1.0000**	0.8114	**0.8243**
Wine	**1.0000**	**1.0000**	**0.9608**	**0.9608**
Flags	**0.9908**	0.9897	0.5771	**0.6500**
Autos	**0.9680**	0.9664	0.7655	**0.7907**
Images	**0.9788**	0.9783	0.8714	**0.8905**
Soybean	**0.9974**	**0.9974**	0.8286	**0.9253**
Vehicle	**0.8559**	0.8551	0.6656	**0.6714**
Tic	**1.0000**	**1.0000**	0.9238	**0.9917**
German	**0.0249**	**0.0249**	**0.7030**	**0.7030**
Anneal	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Bumps	**0.4644**	**0.4644**	**0.9156**	0.9133
BHP	**0.9962**	**0.9962**	0.9953	**0.9962**
DRD	**0.2161**	**0.2161**	**0.6377**	**0.6377**
Mushroom	**0.5623**	0.4852	0.4840	**0.6273**
PB	**0.9630**	0.9582	0.9613	**0.9616**
Mean	**0.8561**	0.8519	0.8375	**0.8625**

Table 6
Classification accuracy comparison.

Datasets	HSRM-R	HRS	HE	FS	LS
Hep.	**0.9163**	0.8713	0.8654	0.8838	0.8708
Iono	**0.9288**	0.8917	0.8745	0.9087	0.8860
Horse	**0.9728**	0.9649	0.9649	0.8995	0.8589
Votes	**0.9679**	0.9566	0.9587	0.9449	0.9516
Credit	**0.8333**	0.8087	0.8087	0.8101	0.7217
Zoo	0.9500	0.9500	0.9500	0.8909	**0.9600**
Lym.	**0.8243**	0.8043	0.7976	0.7971	0.7514
Wine	**0.9608**	0.9382	0.9212	0.9497	0.9438
Flags	**0.6500**	0.5832	0.5879	0.6089	0.6079
Autos	**0.7907**	0.7171	0.7121	0.7426	0.6695
Images	**0.8905**	0.8476	0.8571	0.8667	0.8381
Soybean	**0.9253**	0.8828	0.8654	0.8433	0.8434
Vehicle	**0.6714**	0.6655	0.6632	0.6513	0.6312
Tic	**0.9917**	0.8966	0.8883	0.7368	0.7118
German	**0.7030**	**0.7030**	**0.7030**	**0.7030**	**0.7030**
Anneal	**1.0000**	**1.0000**	**1.0000**	0.9978	**1.0000**
Bumps	0.9133	0.9145	0.9149	0.9218	**0.9234**
BHP	**0.9962**	0.9755	0.9568	0.9633	0.9559
DRD	**0.6377**	0.6351	0.6255	0.6255	0.5995
Mushroom	**0.6273**	0.5865	0.5167	0.6233	0.6264
PB	**0.9616**	0.9585	0.9611	0.9607	0.9543
Mean	**0.8625**	0.8358	0.8282	0.8252	0.8099

Table 7
Results of Friedman test.

Algorithm	*p*-value	Hypothesis
HRS	1.62e−4	Rejected
HE	1.62e−4	Rejected
FS	5.70e−5	Rejected
LS	5.79e−4	Rejected

To further enhance the performance analysis, this paper used the method similar to literature [64] and conducted the Friedman test [65] for the classification accuracy. The null hypothesis assumes that HSRM-R algorithm has the same average classification accuracy as the other four algorithms and that the observed differences are random. The classification accuracy of the five algorithms were ranked from 1 to 5 for each dataset with the rank 5 representing the highest classification accuracy. The average ranking of each algorithm, sorting in descending order, was: HSRM-R (4.5476), HRS (2.9762), FS (2.7857), HE (2.5714) and LS (2.1190). According to relevant theories in literature [65], the result of Friedman test was $\chi^2 = 31.33$ and $p - value = 2.62e{-}6$, so the null hypothesis is rejected at the significance level of 0.01. Based on this evidence, this paper carried out a post-hoc analysis by means of a pairwise comparison to identify the differences between the algorithms. HSRM-R algorithm was used as control method and compared with the other algorithms respectively to detect whether HSRM_R algorithm outperforms any other algorithm used in experiment. From Table 7, it is observed that HSRM-R algorithm statistically outperforms all the other algorithms (when considering $\alpha = 0.01$). Readers can read literature [65] for more about Friedman test.

To verify the effect of complexity control on the generalization ability, we compared the number of rules and support coefficient of five algorithms. Results are shown in Table 8. In Table 8, HSRM-R algorithm obtains fewest rules in 12 datasets while HRS, HE, FS and LS algorithm obtains fewest rules in 2 datasets, 2 datasets, 4 datasets and 7 datasets respectively. Because FS and LS algorithm obtains much fewer rules than HSRM-R algorithm in some datasets, the mean number of rules obtains by FS algorithm and LS algorithm is smaller than HSRM-R algorithm. But HSRM-R algorithm obtains fewest rules in most datasets used for experiments. Thus, HSRM-R obtains fewest rules among the five algorithms. The number of rules characterizes the complexity. Thus, HSRM-R controls the complexity well. Meanwhile, it is observed that HSRM-R algorithm obtains the largest support coefficient in 15 datasets and its mean support coefficient is also the largest. The support coefficient represents the strength of rules. Thus, HSRM-R algorithm obtains the strongest rules. This explains why it obtains better classification accuracy than the other four algorithms to some extent.

6.3. Time complexity analysis

After comparison of classification accuracy, this paper also compared the time complexity of the HSRM-R algorithm, HRS algorithm, HE algorithm, FS algorithm and LS algorithm. The average time of 10-fold cross-validation experiment that the five algorithms consume is listed in Table 9.

In Table 9, the time complexity of LS algorithm and FS algorithm is the smallest. The main reason is that the two algorithms only compute the attribute significance once. They compute the scores of all attributes, rank scores in descending order and select top-m attributes. They do not need iteration. By contrast, HSRM-R algorithm, HRS algorithm and HE algorithm compute the attribute significance each iteration until stopping criterion is met. Although LS algorithm and FS algorithm are fast and efficient, they do not consider the combination of attributes or feature redundancy [61,62]. Thus, their classification accuracy in Table 6 is relatively low. Fig. 3 shows the change of classification accuracy with the number of selected attributes in HSRM-R, FS and LS algorithm. The classification accuracy of FS and LS algorithm is lower than HSRM-R algorithm when selecting identical number of attributes. As the number of selected attributes increases, the classification accuracy of LS and FS algorithm first increases, then remains almost unchanged or decreases and then starts to increase again. This shows that LS and FS usually select some redundant attributes. Thus, FS algorithm and LS algorithm usually consume less time but obtain lower classification accuracy than heuristic algorithms.

HSRM-R algorithm consumes much more time than the other algorithms in Table 9. To study the scalability of it, the scatter diagram between its running time and the running time of HRS algorithm is shown in Fig. 4. The running time of mushroom dataset and PB datasets is much bigger than the other datasets. Considering the display effect of the scatter diagram, the two datasets are not shown in Fig. 4. Fig. 4 shows strong linear

Table 8
Comparison of rule number and support coefficient.

Datasets	Rule number					Support coefficient				
	HSRM-R	HRS	HE	FS	LS	HSRM-R	HRS	HE	FS	LS
Hep.	**13.2**	16.8	15.7	17.2	16.7	**0.1544**	0.1170	0.1027	0.1047	0.1039
Iono	**27.1**	38.0	38.6	34.5	39.8	**0.0744**	0.0434	0.0416	0.0575	0.0491
Horse	**13.1**	22.2	22.2	40.9	57.9	**0.1445**	0.0715	0.0715	0.0575	0.0358
Votes	17.1	17.5	**15.6**	18.9	18.7	**0.1368**	0.1178	0.1281	0.0999	0.0978
Credit	107.4	126.7	126.0	**100.2**	144.9	**0.0212**	0.0161	0.0161	0.0175	0.0103
Zoo	**9.2**	10.3	10.0	9.7	10.6	**0.1119**	0.0993	0.1019	0.1033	0.0971
Lym.	**25.1**	39.0	39.1	31.7	40.6	**0.0631**	0.0385	0.0373	0.0504	0.0366
Wine	**9.3**	12.5	12.3	11.8	11.3	**0.1610**	0.1005	0.1004	0.1457	0.1433
Flags	**56.0**	73.5	73.4	63.4	66.1	**0.0243**	0.0174	0.0167	0.0211	0.0200
Autos	48.6	54.8	54.3	50.2	**38.0**	0.0259	0.0222	0.0207	0.0250	**0.0293**
Images	**22.8**	27.4	25.8	23.4	23.8	**0.0501**	0.0405	0.0399	0.0458	0.0453
Soybean	59.2	86.9	87.4	33.7	**26.4**	0.0205	0.0141	0.0143	0.0307	**0.0397**
Vehicle	172.4	185.7	180.1	99.9	**90.8**	0.0091	0.0083	0.0081	0.0116	**0.0129**
Tic	**23.7**	120.2	119.0	103.6	101.2	**0.0445**	0.0125	0.0126	0.0132	0.0137
German	**13.2**	**13.2**	**13.2**	**13.2**	**13.2**	**0.1944**	**0.1944**	**0.1944**	**0.1944**	**0.1944**
Anneal	**5.0**	**5.0**	6.0	20.2	**5.0**	**0.2000**	**0.2000**	0.1705	0.0585	**0.2000**
Bumps	148.9	149.1	154.3	91.0	**81.8**	0.0090	0.0090	0.0084	**0.0142**	**0.0142**
BHP	**46.8**	60.4	65.2	57.6	56.8	**0.0286**	0.0210	0.0173	0.0214	0.0218
DRD	72.0	72.0	67.2	29.4	**21.7**	0.0273	0.0273	0.0292	0.0481	**0.0644**
Mushroom	214.2	776.0	948.4	**25.0**	28.8	0.0196	0.0015	0.0014	**0.0429**	0.0354
PB	201.0	204.7	208	**170.2**	191.5	**0.0152**	0.0105	0.0089	0.0079	0.0082
Mean	62.2	100.6	108.7	**49.8**	51.7	**0.0731**	0.0563	0.0544	0.0558	0.0606

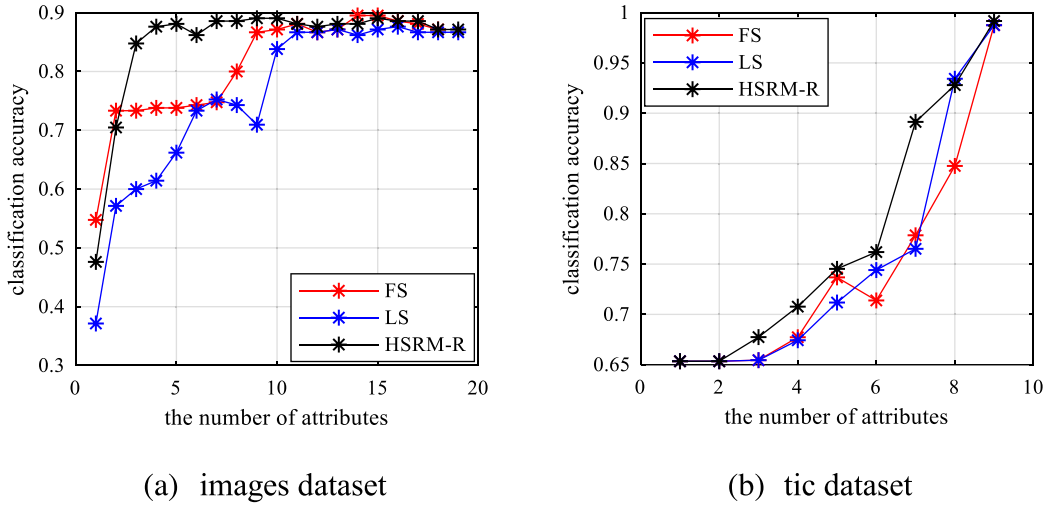


Fig. 3. The change of classification accuracy with the number of attributes.

relationship. Thus, this paper performed linear regression analysis [66] for all the datasets except mushroom and PB to verify the linear relationship. The null hypothesis of linear regression analysis assumes that there is no linear regression. R Square, adjusted R Square and p-value are computed. Usually, R Square and adjusted R Square are elements of [0,1]. Larger R Square and adjusted R Square represent stronger linear relationship in linear regression. The result of linear regression is that $T_{HSRM-R} = 10.095T_{HRS} + 21.193$ with R Square = 0.8621, adjusted R Square = 0.8540 and p-value = 9.925e−9, where T_{HSRM-R} and T_{HRS} denotes the running time of HSRM-R algorithm and HRS algorithm respectively. Thus, the null hypothesis is rejected at the significance level of 0.01. For the two datasets that was not considered in linear regression, namely PB dataset and mushroom dataset, T_{HSRM-R} is 12.65 and 4.54 times of T_{HRS} respectively. Therefore, T_{HSRM-R} is usually about 10 times of T_{HRS} and sometimes T_{HSRM-R} may be less than 10 times of T_{HRS}.

To explain the results of linear regression, some theoretical analysis is performed. HSRM-R algorithm, HRS algorithm and HE algorithm are all heuristic algorithms. They select the attribute with the maximal significance each iteration until the stopping

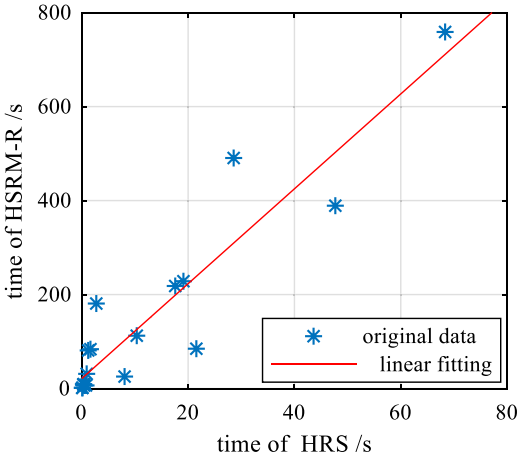


Fig. 4. Scatter diagram of T_{HSRM-R} and T_{HRS}.

Table 9
Running time (unit: second).

Datasets	HSRM-R	HRS	HE	FS	LS
Hep.	6.40	0.46	0.43	0.18	0.12
Iono	83.70	1.63	1.45	0.78	0.67
Horse	9.03	0.37	0.47	0.63	0.78
Votes	6.59	0.80	0.68	0.16	0.14
Credit	228.34	19.12	22.07	3.61	5.42
Zoo	1.52	0.12	0.21	0.10	0.03
Lym.	9.00	0.30	0.43	0.26	0.18
Wine	1.42	0.10	0.18	0.11	0.05
Flags	81.08	1.21	0.97	0.45	0.42
Autos	31.22	0.92	0.82	0.31	0.16
Images	8.30	0.34	0.30	0.16	0.08
Soybean	490.56	28.57	17.59	0.53	0.36
Vehicle	389.07	47.69	43.13	1.33	1.08
Tic	84.74	21.55	26.34	1.60	1.63
German	25.47	8.05	3.73	0.40	0.35
Anneal	180.89	2.73	3.97	0.59	0.94
Bumps	758.79	68.34	74.88	3.95	3.46
BHP	218.20	17.58	14.45	1.43	1.36
DRD	112.43	10.34	12.83	0.49	0.41
Mushroom	39695.29	8747.35	36013.52	48.14	63.55
PB	2741.97	216.81	220.77	15.43	33.46
Mean	2150.67	437.83	1736.15	3.84	5.46

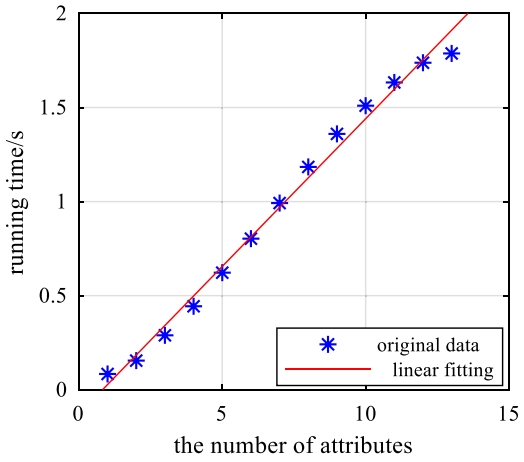


Fig. 5. Scatter diagram of wine dataset.

criterion is satisfied. If there are n attributes in the conditional attribute set and m attributes are selected ($m \leq n$), then they need to compute the significance of $n-m+1$ attributes in the mth iteration. Let T_0, T_1 and T_2 denote the time for computing the dependency degree, the time for extracting rules and the time for computing the information entropy respectively. It is assumed that T_0 and T_1 are constants in each iteration. If the three algorithms both select m attributes, then the consumed time of HSRM-R algorithm, HRS algorithm and HE algorithm are $(T_0+T_1)\sum_{i=1}^{m}(n-i+1)$, $T_0\sum_{i=1}^{m}(n-i+1)$ and $T_2\sum_{i=1}^{m}(n-i+1)$ respectively. Usually T_2 are close to T_0. Thus, HE and HRS algorithm consume similar time. For HSRM-R algorithm, its running time is T_1/T_0 times more than HRS algorithm. If HSRM-R algorithm selects m_1 more attributes than HRS algorithm ($m_1 \leq n-m$), then HSRM-R algorithm needs to consume $(T_0+T_1)\sum_{i=m+1}^{m_1+m}(n-i+1)$ more time, which is also a finite number. Above analysis are consistent with the above linear regression results. Therefore, the time complexity of HSRM-R algorithm is finite, and this algorithm is scalable.

Besides, this paper also studies the relationship between the running time and the number of selected attributes. The scatter in wine dataset is shown in Fig. 5, which shows strong linear relationship. To study whether the relationship also exists in the other datasets, linear regression is performed. Table 10 shows the results of linear regression in all datasets. In Table 10, the p-value is much less than 0.01 in all datasets, so the null hypothesis is rejected at the significance level of 0.01. Meanwhile, R Square and adjusted R Square exceed 0.95 in all datasets except tic, mushroom and PB dataset. Thus, there is strong linear relationship between running time and the number of selected attributes. This further shows the scalability of HSRM-R algorithm.

Based on the above analysis, the time complexity of HSRM-R algorithm increases linearly with the number of selected attributes. Therefore, to further ensure its scalability in relatively large datasets, its stopping criterion can be modified to select fewer attributes. To avoid deleting important attributes, this paper studied the change of attribute significance with the number of selected attributes. The trends of two typical datasets are shown in Fig. 6. In Fig. 6, the significance of attributes begins to decrease when the number of selected attributes increases to a certain value. With the further increase of selected attributes, the attribute significance decreases to a small value. Thus, the HSRM-R algorithm can be stopped at this time to improve its computation efficiency. Formally, the improved HSRM-R algorithm stops when $SIG_{stru}(a,B,D) \leq \delta$ instead of $SIG_{stru}(a,B,D) < 0$, where $\delta \in [0,1]$. Usually, δ is close to zero. In our experiment, $\delta = 0.01$. The comparison between original HSRM-R algorithm and improved HSRM-R algorithm is shown in Table 11. Compared to the original HSRM-R algorithm, the improved HSRM-R algorithm selects fewer attributes, consumes much less time while its classification accuracy remains almost unchanged or increases a little. Therefore, the improvement can effectively ensure the scalability of HSRM-R algorithm.

In summary, the time complexity of HSRM-R algorithm is usually about 10 times of HRS algorithm and its time complexity increases linearly with the number of selected attributes. Its efficiency can be significantly improved by introducing a threshold when applied to large datasets. Although the time complexity of HSRM-R algorithm is the highest in experiments, it obtains the highest classification accuracy. HSRM-R algorithm provides an approach to guaranteeing the generalization ability theoretically in the case where users require high classification accuracy.

7. Discussion

To improve the classification ability in unseen objects, this paper introduced the SRM principle into attribute reduction, defined a novel measure of attribute significance and proposed the HSRM-R algorithm based on the new attribute significance. By selecting proper complexity coefficient w, the complexity decreases much while the empirical risk almost remains unchanged. Thus, the proposed algorithm can realize the structural risk minimization of the rough set-based classifier. Meanwhile, the support coefficient is also large, which means the extracted rules can represent the intrinsic characteristics well.

In the proposed HSRM-R algorithm, the complexity weight w is an important parameter to adjust the relative proportion of complexity and empirical risk in the proposed HSRM-R algorithm. The selection of complexity weight w is a tradeoff between the empirical risk and the complexity. As w increases, the complexity decreases while the empirical risk first almost remains unchanged then increases. A moderate weight coefficient can control the complexity effectively on the precondition of relatively small empirical risk, realize the structural risk minimization of rough set-based classifier and obtain good generalization ability. According to experimental results, w should be between 0.5 and 1 in the datasets used in the experiment. The optimal weight coefficient may vary with datasets. For other datasets, the optimal w can be selected through 10-fold cross-validation experiment.

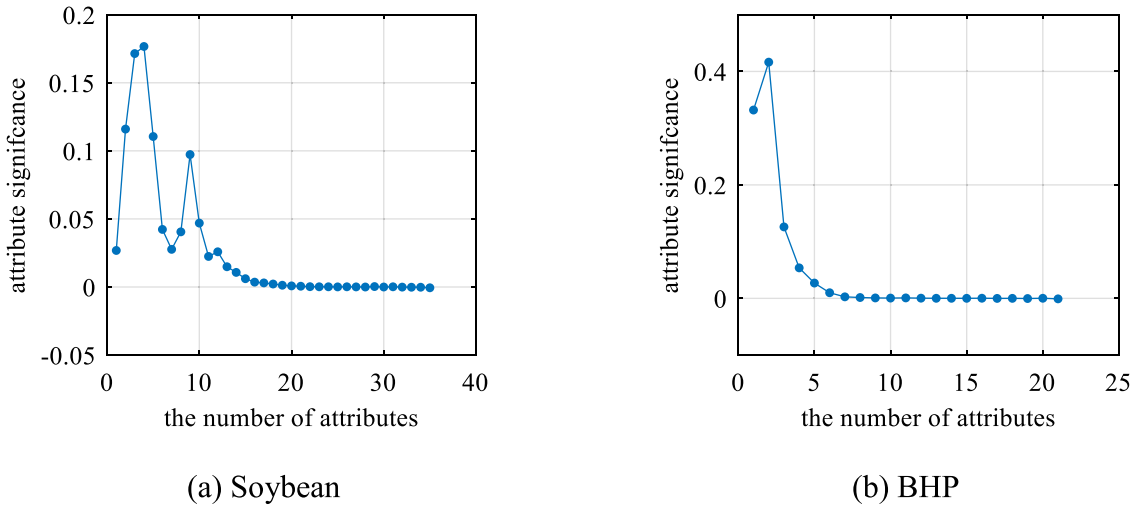


Fig. 6. The relationship between attribute significance and the number of selected attributes.

Table 10
Results of linear regression.

Datasets	Hep.	Iono	Horse	Votes	Credit	Zoo	Lym.	Wine	Flags	Autos
R Square	0.965	0.986	0.990	0.975	0.963	0.990	0.985	0.991	0.992	0.985
Adjusted R Square	0.963	0.985	0.990	0.973	0.961	0.989	0.984	0.990	0.992	0.984
p-value	7.5e−14	5.2e−31	1.4e−21	1.2e−12	1.0e−10	2.5e−15	4.5e−16	1.6e−12	4.9e−29	1.2e−20
Images	Soybean	Vehicle	Tic	German	Anneal	Bumps	BHP	DRD	Mushroom	PB
0.992	0.974	0.939	0.871	0.993	0.987	0.955	0.980	0.964	0.727	0.929
0.991	0.974	0.935	0.853	0.992	0.987	0.952	0.979	0.962	0.713	0.920
3.1e−19	8.2e−28	3.8e−11	2.3e−4	6.2e−25	1.6e−35	3.6e−12	1.2e−17	9.9e−14	4.7e−7	7.2e−6

Table 11
Comparison between improved HSRM-R and original HSRM-R.

Datasets		soybean	anneal	bumps	BHP	DRD	Mushroom	PB
Time/s	Original	490.56	180.09	758.79	218.20	112.43	39695.29	2741.97
	Improved	173.79	7.61	257.72	83.96	54.96	3884.24	1614.99
Attribute number	Original	30.6	35.2	17.9	17.4	19.0	18.1	9.9
	Improved	15.3	5.0	7.2	7.4	12.5	8.0	8.4
Accuracy	Original	0.9253	1.0000	0.9133	0.9962	0.6377	0.6273	0.9616
	Improved	0.9210	1.0000	0.9299	0.9953	0.6359	0.6416	0.9625

The conventional HRS algorithm also controls the complexity to some extent by introducing the threshold. But it controls the complexity merely by decreasing the dependency degree. Thus, the decrease of complexity is at the cost of the increase of the empirical risk. When the complexity decreases, the empirical risk also increases. By contrast, the biggest advantage of the proposed HSRM-R algorithm is that it can decrease the complexity while remaining the empirical risk almost unchanged. Additionally, Fisher score and Laplacian score do not consider redundant attributes or the combination of attributes and they does not control the complexity. Thus, the proposed HSRM-R algorithm can find a better tradeoff between the empirical risk and the complexity and obtain smaller structural risk. Therefore, the proposed HSRM-R algorithm can obtain better generalization ability.

8. Conclusion

To improve the generalization ability of rough set-based classifier, this paper introduces structural risk minimization (SRM) inductive principle into attribute reduction and proposes that the number of rules can characterize the actual complexity of rough set-based classifier effectively. By introducing the complexity weight to find a tradeoff between the empirical risk and complexity, this paper defined a novel measure of attribute significance with complexity weight. Based on the new attribute significance, this paper developed a new heuristic attribute reduction algorithm called HSRM-R algorithm. Three main conclusions are drawn from this study.

Firstly, HSRM-R algorithm controls the complexity directly, decreases the complexity while remaining the empirical risk almost unchanged and obtains small structural risk. The 10-fold cross-validation experiments in 21 UCI datasets show that the proposed HSRM-R algorithm obtains better generalization ability than conventional HRS algorithm. Further experiment shows that HSRM-R algorithm obtains larger dependency degree, fewer rules and larger support coefficient than conventional HRS algorithm. Additionally, experiment also shows that HSRM-R algorithm can also obtain better classification accuracy than HRS algorithm when they have similar dependency degree. This further proves that introducing complexity control can effectively improve the generalization ability.

Secondly, experiments show that the proposed HSRM-R algorithm obtains better classification accuracy than other famous attribute reduction algorithms including heuristic entropy-based algorithm, Fisher score and Laplacian score. Statistical test is also performed to verify its superiorities in generalization ability. Further experiments show that HSRM-R algorithm can obtain

fewer rules and larger support coefficient than other conventional algorithms. This shows that HSRM-R can control the complexity well and extract strong rules, which can explain why it obtains better generalization ability to some extent.

Thirdly, the proposed HSRM-R algorithm has the highest time complexity while Fisher score and Laplacian score have the lowest time complexity in the experiments. But Fisher score and Laplacian score do not consider the redundancy of attributes, thus their classification accuracy is usually relatively low. The time complexity of HSRM-R algorithm is usually about 10 times of HRS algorithm in experiments and its time complexity increases linearly with the number of selected attributes. Its efficiency can be significantly improved by introducing a threshold when applied to relatively large datasets. Although the time complexity of HSRM-R algorithm is the highest in experiments, it obtains the highest classification accuracy and it is scalable. The proposed HSRM-R algorithm provides an approach to guaranteeing the generalization ability theoretically in the case where users require high classification accuracy.

Acknowledgments

This work was supported by National Key R&D Program of China No. 2017YFB0902100 and National Science and Technology Major Project of China No. 2017-I-0007-0008. The authors would like to thank the anonymous reviewers for their careful reading of the paper and valuable suggestions to refine this work.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.105543.

References

[1] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[2] J. Dai, Q. Xu, Approximations and uncertainty measures in incomplete information systems, Inf. Sci. (Ny). 198 (2012) 62–80.
[3] K.Y. Huang, I.H. Li, A multi-attribute decision-making model for the robust classification of multiple inputs and outputs datasets with uncertainty, Appl. Soft Comput. 38 (2016) 176–189, http://dx.doi.org/10.1016/j.asoc.2015.09.015.
[4] J. Shi, Y. Lei, Y. Zhou, M. Gong, Enhanced rough fuzzy c-means algorithm with strict rough sets properties, Appl. Soft Comput. 46 (2016) 827–850, http://dx.doi.org/10.1016/j.asoc.2015.12.031.
[5] J. Zhan, M.I. Ali, N. Mehmood, On a novel uncertain soft set model: Z-soft fuzzy rough set model and corresponding decision making methods, Appl. Soft Comput. 56 (2017) 446–457, http://dx.doi.org/10.1016/j.asoc.2017.03.038.
[6] Z. Pawlak, On conflicts, Int. J. Man. Mach. Stud. 21 (1984) 127–134.
[7] Y. Liu, Y. Lin, Intuitionistic fuzzy rough set model based on conflict distance and applications, Appl. Soft Comput. 31 (2015) 266–273, http://dx.doi.org/10.1016/j.asoc.2015.02.045.
[8] B. Sun, W. Ma, H. Zhao, Rough set-based conflict analysis model and method over two universes, Inf. Sci. (Ny). 372 (2016) 111–125, http://dx.doi.org/10.1016/j.ins.2016.08.030.
[9] R.T. Das, K.K. Ang, C. Quek, ieRSPOP: A novel incremental rough set-based pseudo outer-product with ensemble learning, Appl. Soft Comput. 46 (2016) 170–186, http://dx.doi.org/10.1016/j.asoc.2016.04.015.
[10] X. Xie, X. Qin, C. Yu, X. Xu, Test-cost-sensitive rough set based approach for minimum weight vertex cover problem, Appl. Soft Comput. 64 (2018) 423–435, http://dx.doi.org/10.1016/j.asoc.2017.12.023.
[11] Y.-C. Hu, Flow-based tolerance rough sets for pattern classification, Appl. Soft Comput. 27 (2015) 322–331, http://dx.doi.org/10.1016/j.asoc.2014.11.021.
[12] K.Y. Huang, An enhanced classification method comprising a genetic algorithm, rough set theory and a modified PBMF-index function, Appl. Soft Comput. 12 (2012) 46–63, http://dx.doi.org/10.1016/j.asoc.2011.09.009.
[13] Y. Hong-Wei, T. Xindi, Based on rough sets and L1 regularization of the fault diagnosis of linear regression model, in: 2016 Int. Conf. Intell. Transp. Big Data Smart City, 2016, pp. 490–492, http://dx.doi.org/10.1109/icitbs.2016.145.
[14] Y.E. Shao, C.-D. Hou, C.-C. Chiu, Hybrid intelligent modeling schemes for heart disease classification, Appl. Soft Comput. 14 (2014) 47–52, http://dx.doi.org/10.1016/j.asor.2013.09.020.
[15] Y. Kaya, M. Uyar, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, Appl. Soft Comput. 13 (2013) 3429–3438, http://dx.doi.org/10.1016/j.asoc.2013.03.008.
[16] F. Wang, J. Liang, C. Dang, Attribute reduction for dynamic data sets, Appl. Soft Comput. 13 (2013) 676–689.
[17] A. Phophalia, S.K. Mitra, Rough set based bilateral filter design for denoising brain MR images, Appl. Soft Comput. 33 (2015) 1–14.
[18] D. Liang, W. Pedrycz, D. Liu, P. Hu, Three-way decisions based on decision-theoretic rough sets under linguistic assessment with the aid of group decision making, Appl. Soft Comput. 29 (2015) 256–269, http://dx.doi.org/10.1016/j.asoc.2015.01.008.
[19] K.R. Singh, M.A. Zaveri, M.M. Raghuwanshi, Rough membership function based illumination classifier for illumination invariant face recognition, Appl. Soft Comput. 13 (2013) 4105–4117, http://dx.doi.org/10.1016/j.asoc.2013.04.012.
[20] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: A review, Appl. Soft Comput. J. 9 (2009) 1–12.
[21] X. Hu, N. Cercone, Learning in relational databases: A rough set approach, Comput. Intell. 11 (1995) 323.
[22] T. Jing, W. Quan, Y. Bing, Y. Dan, A rough set algorithm for attribute reduction via mutual information and conditional entropy, in: 2013 10th Int. Conf. Fuzzy Syst. Knowl. Discov., 2013, pp. 567–571, http://dx.doi.org/10.1109/FSKD.2013.6816261.
[23] H. Sun, R. Wang, B. Xie, Y. Tian, Continuous attribute reduction method based on an automatic clustering algorithm and decision entropy, in: Chinese Control Conf., 2017.
[24] T. Yan, C. Han, Entropy based attribute reduction approach for incomplete decision table, in: Int. Conf. Inf. Fusion, 2017, pp. 1–8.
[25] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[26] Q. Hu, H. Zhao, Z. Xie, D. Yu, Consistency based attribute reduction, in: Adv. Knowl. Discov. Data Mining, Pacific-Asia Conf. PAKDD 2007, Nanjing, China, May 22–25, 2007, Proc., 2007, pp. 96–107.
[27] J. Liu, Q. Hu, D. Yu, Weighted rough set learning: towards a subjective approach, in: Pacific-Asia Conf. Adv. Knowl. Discov. Data Min., 2007, pp. 696–703.
[28] W. Ziarko, Variable precision rough set model, J. Comput. System Sci. 46 (1993) 39–59, http://dx.doi.org/10.1016/0022-0000(93)90048-2.
[29] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, Theory Decis. Libr. 11 (1992) 331–362.
[30] S.H. Teng, M. Lu, A.F. Yang, J. Zhang, Y.J. Nian, M. He, Efficient attribute reduction from the viewpoint of discernibility, Inf. Sci. (Ny). 326 (2016) 297–314, http://dx.doi.org/10.1016/j.ins.2015.07.052.
[31] Y.Y. Yao, Y. Zhao, Discernibility matrix simplification for constructing attribute reducts, Inf. Sci. (Ny). 179 (2009) 867–882, http://dx.doi.org/10.1016/j.ins.2008.11.020.
[32] A.R. Hedar, M.A. Omar, A.A. Sewisy, Rough sets attribute reduction using an accelerated genetic algorithm, in: Ieee/acis Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/distributed Comput., 2015, pp. 1–7.
[33] G. Dai, Z. Wang, C. Yang, H. Liu, A multi-granularity rough set algorithm for attribute reduction through particles particle swarm optimization, in: Int. Comput. Eng. Conf., 2015, pp. 303–307.
[34] L. Wang, L. Ma, Q. Bian, X. Zhao, Rough set attributes reduction based on adaptive PBIL algorithm, in: IEEE Int. Conf. Inf. Theory Inf. Secur., 2010, pp. 21–24.
[35] F. Min, X. Du, H. Qiu, Q. Liu, Minimal attribute space bias for attribute reduction, in: Rough Sets Knowl. Technol. Second Int. Conf. RSKT 2007, Toronto, Canada, May 14-16, 2007, Proc., 2007, pp. 379–386.
[36] X. Jia, L. Shang, B. Zhou, Y. Yao, Generalized attribute reduct in rough set theory, Knowl.-Based Syst. 91 (2016) 204–218, http://dx.doi.org/10.1016/j.knosys.2015.05.017.
[37] V.N. Vapnik, A.J. Chervonenkis, Theory of pattern recognition, 1974.
[38] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.
[39] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (1999) 988–999.
[40] D. Liu, H. Qian, G. Dai, Z. Zhang, An iterative SVM approach to feature selection and classification in high-dimensional datasets, Pattern Recognit. 46 (2013) 2531–2537, http://dx.doi.org/10.1016/j.patcog.2013.02.007.
[41] J. Luo, C. Wei, H. Dai, J. Yuan, Robust LS-SVM-based adaptive constrained control for a class of uncertain nonlinear systems with time-varying predefined performance, Commun. Nonlinear Sci. Numer. Simul. 56 (2018) 561–587, http://dx.doi.org/10.1016/j.cnsns.2017.09.004.
[42] Z.-M. Wang, N. Han, Z.-M. Yuan, Z.-H. Wu, Feature selection for high-dimensional data based on ridge regression and SVM and its application in peptide QSAR modeling, Acta Physico-Chimica Sin. 29 (2013) 498–507, http://dx.doi.org/10.3866/pku.whxb201301042.

[43] J. Nong, The design of RBF neural networks and experimentation for solving overfitting problem, in: Int. Conf. Electron. Optoelectron., 2011, pp. V1-75-V1-78.
[44] D.A.G. Vieira, J.A. Vasconcelos, R.R. Saldanha, Recent advances in neural networks structural risk minimization based on multiobjective complexity control algorithms, in: Mach. Learn., InTech, 2010.
[45] D. Kim, Minimizing structural risk on decision tree classification, in: Multi-Objective Mach. Learn, Springer, 2006, pp. 241–260.
[46] O.T. Yildiz, VC-dimension of univariate decision trees, IEEE Trans. Neural Networks Learn. Syst. 26 (2015) 378.
[47] Z. Pawlak, A. Skowron, Rudiments of rough sets, Inf. Sci. (Ny). 177 (2007) 3–27.
[48] 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.
[49] D. Zak, Approximate entropy reducts, Fund. Inform. 53 (2002) 365–390.
[50] J. Liang, K.S. Chin, C. Dang, R.C.M. Yam, A new method for measuring uncertainty and fuzziness in rough set theory, Int. J. Gen. Syst. 31 (2002) 331–342.
[51] J. Liang, Z. Xu, The algorithm on knowledge reduction in incomplete information systems, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (2002) 95–103.
[52] Y. Qian, J. Liang, Combination entropy and combination granulation in rough set theory, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 16 (2008) 179–193.
[53] Y. Yao, Y. Zhao, J. Wang, On reduct construction algorithms, Lecture Notes in Comput. Sci. 5150 (2008) 100–117, (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics, LNCS), http://dx.doi.org/10.1007/978-3-540-87563-5_6.
[54] J.-S. Mi, W.-Z. Wu, W.-X. Zhang, Approaches to knowledge reduction based on variable precision rough set model, Inf. Sci. (Ny). 159 (2004) 255–272.
[55] D. Chen, Y. Yang, Z. Dong, An incremental algorithm for attribute reduction with variable precision rough sets, Appl. Soft Comput. 45 (2016) 129–149, http://dx.doi.org/10.1016/j.asoc.2016.04.003.
[56] J. Zhou, D. Miao, Q. Feng, L. Sun, Research on complete algorithms for minimal attribute reduction, in: Rough Sets Knowl. Technol. Int. Conf. Rskt 2009, Gold Coast, Aust. July 14–16, 2009, Proc., 2009, pp. 152–159.
[57] C. Xu, F. Min, Weighted reduction for decision tables, in: Int. Conf. Fuzzy Syst. Knowl. Discov, Springer, 2006, pp. 246–255.
[58] J.W. Grzymala-Busse, LERS-a system for learning from examples based on rough sets, in: Intell. Decis. Support, Springer, 1992, pp. 3–18.
[59] Asuncion, D. Newman, Uci Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2007, p. 1994, (n.d.).
[60] J.F. Liu, Q.H. Hu, D.R. Yu, A weighted rough set based method developed for class imbalance learning, Inf. Sci. (Ny). 178 (2008) 1235–1256, http://dx.doi.org/10.1016/j.ins.2007.10.002.
[61] Q. Gu, Z. Li, J. Han, Generalized fisher score for feature selection, 2012, arXiv Prepr. arXiv:1202.3725.
[62] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, Data Classif. Algorithms Appl. (2014) 37.
[63] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. (2006) 507–514.
[64] E. Queiroga, A. Subramanian, L. dos Anjos F. Cabral, Continuous greedy randomized adaptive search procedure for data clustering, Appl. Soft Comput. 72 (2018) 43–55, http://dx.doi.org/10.1016/j.asoc.2018.07.031.
[65] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Amer. Statist. Assoc. 32 (1937) 675–701, http://dx.doi.org/10.1080/01621459.1937.10503522.
[66] G.A.F. Seber, A.J. Lee, Linear Regression Analysis, John Wiley & Sons, 2012.

Jinfu Liu received his Ph.D. degree in Harbin Institute of Technology. He is currently an associate professor in School of Energy Science and Engineering of Harbin Institute of Technology. His research interests include fault diagnosis, rough sets, machine learning and data mining.

Mingliang Bai is currently a master candidate in School of Energy Science and Engineering of Harbin Institute of Technology. His research interests include rough sets, machine learning and fault diagnosis.

Na Jiang is currently a master candidate in School of Energy Science and Engineering of Harbin Institute of Technology. Her research interests include rough sets and data mining.

Daren Yu received his Ph.D. degree in 1996 from Harbin Institute of Technology. He is currently a professor in School of Energy Science and Engineering of Harbin Institute of Technology. His research interests include rough sets, machine learning, data mining, fault diagnosis and automation control.