Information Sciences 520 (2020) 373–388

Contents lists available at ScienceDirect

Information Sciences

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

A neighborhood rough set model with nominal metric embedding

Sheng Luo [a,*], Duoqian Miao [b,c,**], Zhifei Zhang [b,c], Yuanjian Zhang [b,c], Shengdan Hu [b,c]

[a] *School of Computer and Information, Shanghai Second Polytechnic University, Shanghai, 201209, China*
[b] *Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China*
[c] *Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 201804, China*

ARTICLE INFO

Article history:
Received 16 March 2018
Revised 4 February 2020
Accepted 6 February 2020
Available online 7 February 2020

MSC:
00-01
99-00

Keywords:
Attribute reduction
Neighborhood rough set
Metric learning
Nominal data

ABSTRACT

Rough set theory is an essential tool for measuring uncertainty, which has been widely applied in attribute reduction algorithms. Most of the related researches focus on how to update the lower and the upper approximation operator to match data characteristics or how to improve the efficiency of the attribute reduction algorithm. However, in the nominal data environment, existing rough set models that use the Hamming metric and its variants to evaluate the relations between nominal objects can not capture the inherent ordered relationships and statistic information from nominal values due to the complexity of data. The missing information will affect the accuracy and validity of the data representation, thereby reducing the reliability of rough set models. To overcome this challenge, we propose a novel object dissimilarity measure, i.e., relative object dissimilarity metric(RODM) that learned from nominal data to replace the Hamming metric and then construct a ψ-neighborhood rough set model. It extends the classical rough set model to a robust, representative, and effective model which is close to the characteristics of nominal data. Based on the ψ-neighborhood rough set model, we propose a heuristic two-stage attribute reduction algorithm(HTSAR) to perform the feature selection task. Experiments show that the ψ-neighborhood rough set model can take advantage of more potential knowledge in nominal data and achieve better performance for attribute reduction than the existing rough set model.

1. Introduction

Feature selection is a crucial concept in machine learning, and it plays an essential role in many areas including pattern recognition, data mining and representation learning, etc. A large number of methods for solving feature selection tasks are proposed for numerical data, for instance, linear discriminant analysis (LDA) [1], principle component analysis (PCA) [2], independent component analysis (ICA) [3] and Autoencoder [4], etc. However, it is rare to find the similar method to do the same work for nominal data. One of the most significant problems is how to measure the distance (or similarity) between nominal attribute values. In other words, it is difficult to measure the relationships between nominal values due

[*] Corresponding author.
[**] Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China.
E-mail addresses: tjluosheng@gmail.com (S. Luo), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ins.2020.02.015
0020-0255/© 2020 Elsevier Inc. All rights reserved.

to the fuzziness existing in nominal data. Therefore, it is a big challenge to perform feature selection in nominal data environments. Rough set theory is a mathematical tool for handling vague and incomplete information, which is first proposed by Pawlak [5]. The rough set model provides a pair of lower and upper approximation operators, which can map data to an approximation space and measure uncertainty existing in the attribute selection problem. In rough set contexts, the feature selection is also called attribute reduction, which target is to search a minimum attribute set to keep the abilities that are held by attribute sets as a whole for classification or other tasks.

Most existing rough set models, which are defined by equivalence relations for nominal data, rely on the comparison result of nominal attribute values. Therefore, the strategy of how to measure the similarity (or distance) of nominal attribute values becomes a crucial point to construct rough set models. In nominal data environments, it is easy to see that the main strategy for rough set models to distinguish indiscernibility relationships is to directly compare two nominal values, returning 1 if the results are the same, and 0 otherwise. Although this method is easy to implement and understand, its disadvantages are obvious due to the loss of information. For example, given a nominal attribute such as “education” with a value domain {“pupil”, “middle school student”, “undergraduate”, “master candidate”, “doctoral candidate”}, it is not difficult to find that the distance between any one pair of nominal values is 0, if nominal values being compared are the same, and 1 otherwise, by using the strategy above. However, according to common sense, the distance between “pupil” and “doctoral candidate” is not the same as the distance between “master candidate” and “doctoral candidate”. The main factor leading to information loss is that it ignores ordered relationships or other potential relationships within nominal values. Therefore, the strategy of directly comparing whether the attribute values are exactly equivalent implies the loss of potential information. How to evaluate the similarity between two nominal objects and maintain more implicit information has become a challenging problem. There are many ways to evaluate the similarity between two objects with numeric features such as the Euclidean distance, the Manhattan distance, and the Mahalanobis distance, etc. However, these measures are hardly applied to the nominal circumstance. In the rough set theory context, the most widely used similarity metric for nominal data is the Hamming distance[6], which means the dissimilarity between the same nominal values is “0”, and “1” otherwise. As is mentioned above, the main drawback of this metric is that it does not take the latent ordered information and statistical properties of nominal values into account. Although the main starting point of rough set theory is to handle inaccurately, incomplete, and inconsistent data, indiscernibility relationships which are defined by the Hamming distance lead to the information loss. Then, the loss makes the rough set model lack the ability to represent nominal data well. What’s more, by employing the Hamming distance, small changes in nominal data will result in the considerable discrepancies of equivalence relations, which is analog to ill-condition phenomena in matrix analysis. In other words, the rough set model equipped with the Hamming distance is sensitive to the changes in nominal data.

To overcome these challenges, in this paper, building on the idea of correlation between nominal attribute values and classification labels, we propose a novel data metric, called relative object dissimilarity metric(RODM). The metric can learn a latent ordered relationship from data by using the statistic information of nominal values. Based on RODM, we propose a new rough set model, named ψ-neighborhood rough set. The model uses RODM to update comparison strategies existing in rough set models. Unlike existing rough set models, the ψ-neighborhood rough set model can expand the indiscernibility relationship into a new dynamic indiscernibility relationship through the scale parameter ψ. The dynamic indiscernibility relationship can make the ψ-neighborhood rough set model more in line with true characteristics of nominal data. Based on the ψ-neighborhood rough set model, we also propose a new attribute reduction algorithm, called heuristic two-stage attribute reduction algorithm (HTSAR), which has the ability to obtain a more representative attribute subset than the existing algorithms. The experimental section proves that our method is effective and superior to the existing rough set models. The contributions of the paper are summarized as follows:

- We propose a new kind of distance measure, i.e., RODM, for nominal data. By using RODM, we can take advantage of the statistic information of nominal data and map the distance between nominal value pairs to a new continuous measuring space.
- We propose a ψ-neighborhood rough set model. Through the setting of the parameter ψ, we have the flexibility to control the extent of indiscernible relationships within nominal objects. The model fully considers the statistical information of nominal data and conforms to the inherent characteristics of the data.
- We propose a heuristic two-stage attribute reduction algorithm, i.e., HTSAR. The algorithm is divided into two parts. The first part of the method maximizes the information gain between different attributes and generates a pool of candidate attributes according to the information gain. In the second part, the algorithm selects the attribute that is most conducive to improving the classification ability from the candidate attributes pool.

The rest of the paper is organized as follows. In Section 2, we introduce some related works. In Section 3, we introduce the preliminary notions related to the Pawlak rough set model. Then, we formulate the research problem and clarify some basic concepts. In Section 4, we introduce the nominal objects similarity measure and define the ψ-neighborhood rough set model. Then, we construct the heuristic two-stage attribute reduction algorithm. In Section 5, we first introduce the data preparation and experiment environment. Then we analyze the performance of HTSAR and the quality of the selected attribute set which generated by HTSAR. Finally, we analyze the impact of parameter ψ in details. In Section 6, we draw a conclusion of this paper.

2. Related work

2.1. Rough set models

Rough set theory is a mathematical tool proposed by Pawlak [5] for dealing with imprecise, incomplete and inconsistent data. The rough set model that takes advantage of equivalence relationships provides a new perspective for classifying nominal objects by employing a pair of lower and upper approximation operators. However, due to the rigor of equivalence relationships, the model cannot deal with the uncertainty of the relationship existing in nominal objects. Many researchers have proposed many methods to update the classic rough set model with various forms. These works can be roughly divided into the following types.

2.1.1. Probabilistic methods

From the probabilistic perspective, Wong and Ziarko [7] firstly expand the classical rough set model to the probabilistic rough set model. The approximation space of probabilistic rough set modes is represented by rough membership functions and rough inclusion. Probabilistic rough set models can be divided into three types: variable precision rough set models [8], decision-theoretic rough set models [9] and Bayesian rough set models [10], which have attracted the attention and in-depth study of many researchers. The main difference among those models is their definitions of probabilistic approximations and the meanings of the required parameters [11]. In addition to probabilistic rough set model, there also exist some rough set models related to probabilistic approaches, e.g.,information-theoretic analysis [12], probabilistic rule induction [13] and other studies [14], etc.

2.1.2. Extended object relationships

Except for probabilistic approaches, some researches try to relax the constraints of equivalence relationships to avoid the deficiency of equivalence relationships. Many rough set models use relaxed relationships instead of equivalence relationships to update classic rough set models, such as tolerance relations [15], general binary relation [16], similarity relations [17], dominance relations [18], etc. Besides, some researchers have proposed fuzzy rough set models [19] by replacing equivalence relations with fuzzy relations. Although these relations avoid the inadequacy of the equivalence relations, they still can not take advantage of complex relationships existing in nominal data. For example, the dominance based rough set model requires the priori knowledge of criteria (attributes with preference-order domains), which is sensitive to noise [20].

2.1.3. Data type adaptation

Since real-valued types cannot directly define equivalence relations, some scholars have proposed some mapping methods to convert real-valued systems into data forms that can be used for classic rough sets, thereby extending data type that can be processed by rough set models. For example, Leung et al. [21] proposed a rough set knowledge discovery framework which is formulated for the analysis of interval-valued information systems converted from real-valued raw decision tables. Sun et al. [22] proposed the interval-valued fuzzy rough set theory by combining the interval-value fuzzy set theory with the traditional rough set theory. Besides, some researchers use kernel methods mapping original numerical data into kernel space and redefine relations between objects in that space. For example, Hu et al. [23] proposed a neighborhood rough set model to handle numerical data, and extended the relationship of object pairs to a kernel space [24].

In general, the data type of objects, existing in information tables, can be divided into three categories, i.e., numeric, nominal and mixture. Then the kind of rough set models is also divided into three categories, correspondingly. In this paper we focus on the rough set model which can be applied to the nominal data. Although the existing nominal rough set mode can take advantage of various relations to represent data, the statistic information and latent ordered information of nominal data are still not to be considered, because the comparison strategy among nominal values is the Hamming distance.

2.2. Attribute reduction

Attribute reduction is a crucial application of rough sets theory. In Pawlak's view [5], the reduction is interpreted as selecting a minimal subset of attributes which keep the positive regions unchanged. According to the definition of reduction and the optimization strategy used, attribute reduction can be roughly divided into the following types.

2.2.1. Methods that focus on how to define the definition of reduction to suit different application environments

For the *fuzzy rough model*, Hu et al. [23] generalized rough set model with neighborhood relations and proposed a neighborhood rough set model based attribute reduction algorithm to deal with numerical data. Wu et al. [25] presented a general framework for the study of rough set approximation operators in fuzzy environment in which both constructive and axiomatic approaches are used. Furthermore, there are some attribute reduction algorithm using fuzzy-rough set model [24]. From *algebra viewpoint and information viewpoint*, Wang et al. [26] made a comparative study on quantitative relationship for attribute reduction, and proved its are equivalent for a consistent decision table. For the *covering rough set model*, Zhu et al. [27] introduced the concept of reducts of covering and had proved that the reduction of a covering is the minimal covering that generates the same covering lower approximation or the same covering upper approximation. For the *probabilistic rough set model*, Yao et al. [28] addressed attribute reduction in decision-theoretic rough set models regarding different classification properties and proposed a framework for attribute reduction in decision-theoretic rough set models. Mi et al. [29] introduced the concepts of β lower distribution reduct and β upper distribution reduct based on variable precision rough sets.

2.2.2. Methods that focus on finding fast reduction algorithms to speed up the efficiency of reduction calculations

For the using of *heuristic information*, considering the issue of candidate attribute selection, Miao et al. [30] proposed a mutual information based strategy to speed up the attribute selection process. Although heuristic reduction methods can take advantage of low computational complexity, the output of these algorithms is an approximate optimization solution. For the using of *discernibility matrix*, Skowron et al. firstly introduced discernibility matrix into attribute reduction algorithms [31]. The discernibility matrix is viewed as a form of knowledge representation that can provide a quick algorithm for finding attribute reduction.

2.2.3. Other methods

There also exists some incremental attribute reduction algorithms for dynamically increasing attribute [32]. Besides, Jia et al. [33] introduced a generalized framework of attribute reduct and summarized most of existing attribute reduct algorithms in combination of data properties and user preference.

In summary, attribute reduct algorithms are sensitive to rough set models. If the model changes, the reduction algorithm will also change accordingly.

2.3. Nominal metrics

Measuring the complex relationship in nominal data is an interesting research topic, which is a challenging problem in the machine learning communities. The Hamming distance [6] is one of the most generally used specific distance metric for objects with nominal values. However, the Hamming distance is very simple and it does not perceive the dependencies between object properties. Recent years, a lot of special distance metrics are proposed such as Ahmad's distance metric (ADM) [34], Hong's distance metric(HDM) [35] and the coulpled nominal distance (CNS) [36], etc. HDM takes more attentions on the combination of two attributes, however, it ignores the differences among multiple attributes. Both ADM and CNS consider the relationship between nominal values pair relative to a third attribute. However, the computation cost plays a bottleneck role in their performance. There are also another kind of nominal metrics. It convert nominal values into numerical vectors, then calculate the distance between these vectors like numerical style, for example, the dummy variables related value embedding (DVE) [37], the coupled data embedding (CDE) [38], heterogeneous support vector machine (HSVM) [39] and multi-granularity neighborhood distance metric learning (MGML) [40] etc.

Nominal metrics provides a new viewpoint to mining the complex relationship in nominal data. Therefore, it is useful to let rough set models take advantage of nominal metric learning which could enhance the representation ability of rough set models for data profoundly.

3. Preliminary

In this section, we will introduce the preliminary notions related to the Pawlak rough set model. Then, we will formulate the problem and clarify some basic concepts.

3.1. Pawlak'S rough set model

Definition 1 (Information System)**.** A four tuple, $\mathcal{T} = \langle U, A, V, f \rangle$, is called an information system, which satisfies the following assumptions,

(1) $U = \{U_i\}_{i=1}^{n}$ is a non-empty set, which elements represent objects in one domain, and n is the total number of objects.
(2) $A = \{a\}_{a=1}^{m}$ represents the feature set of an object, and m is the total number of features.
(3) $V = \{V_a\}_{a=1}^{m}$ is a set of the value domain of features. Especially, if the value domain V_a is of continuous then the value is called the numeric value domain, otherwise, nominal value domain.
(4) $f = \{f_a\}_{a=1}^{m}$ is a set of functions, which element represents a mapping from an object U_i to its feature value V_a, that is, f_a: $U_i \rightarrow V_a$.

Definition 2 (Indiscernibility Relationships)**.** Given a subset $P \subseteq A$, it induces an indiscernibility relationship, which is defined as,

$$IND(P) \triangleq \{(x, y) \in U \times U | f(x, c) = f(y, c), \forall c \in P\} \tag{1}$$

For every $u \in U$, the indiscernibility relationship, $IND(P)$, induces an equivalence class, $[u]_P$, i.e.,

$$[u]_P \triangleq \{v \in U | (u, v) \in IND(P)\} \tag{2}$$

Furthermore, the family of all equivalence classes of $IND(P)$, i.e., the partition induced by P, can be denoted as,

$$\Pi_P \triangleq U/P = \{[u]_P | \forall u \in U\} \tag{3}$$

Table 1
A toy example with nominal attributes.

Obj.	Attr.1	Attr.2	Attr.3	Attr.4	Class
o_1	A	F	L	K	C_1
o_2	B	F	L	K	C_1
o_3	E	S	M	K	C_2
o_4	E	G	M	K	C_2
o_5	D	G	N	K	C_2
o_6	D	S	N	K	C_2

Definition 3 (Approximation Operator)**.** Given an information system, $\mathcal{T} = \langle U, A, V, f\rangle$, $P \subseteq A$, and $X \subseteq U$, then the lower and upper approximation operator of X w.r.t. P, is defined as,

$$\begin{aligned} &\underline{P}(X) = \cup\{[u]_P | [u]_P \subseteq X\} \\ &\overline{P}(X) = \cup\{[u]_P | [u]_P \cap X \neq \emptyset\} \end{aligned} \tag{4}$$

and denoted as $\underline{P}(X), \overline{P}(X)$, respectively.

The lower and upper approximation operator divides the domain U of an information system $\mathcal{T}$ into three parts, i.e., $\underline{P}(X), \overline{P}(X)$ and $\overline{P}(X) - \underline{P}(X)$ w.r.t. P, which are called positive region, negative region, and boundary region of X w.r.t. P, respectively.

Definition 4 (Positive, Negative and Boundary Region)**.** The three regions are defined as follows,

(1) Positive Region: $POS_P(X) = \underline{P}(X)$
(2) Negative Region: $NEG_P(X) = U - \overline{P}(X)$
(3) Boundary Region: $BN_P(X) = \overline{P}(X) - \underline{P}(X)$

If $BN_P(X) = \emptyset$, then X is a definable sets w.r.t. P, otherwise a rough set. In the view of rough set theory, the lower approximation operator over a set X w.r.t. P means that the elements in the $\underline{P}(X)$ certainly belongs to the set X. The upper approximation operator over a set X w.r.t. P means that the elements in the $\overline{P}(X)$ possibly belongs to the set X.

Definition 5 (Approximation Accuracy and Quality)**.** Given an information system $\mathcal{T} = \langle U, A = C \cup D, V, f\rangle$, then the accuracy of an approximation of U/D w.r.t. C is defined as,

$$\alpha_C(D) = \frac{|POS_C(D)|}{\sum_{X \in U/D} |\overline{C}(X)|} \tag{5}$$

where C represents the feature sets of domain objects, and D is the decision attribute or classification attribute. The quality of approximation U/D w.r.t. C is defined as,

$$\gamma_C(D) = \frac{|POS_C(D)|}{|U|} \tag{6}$$

3.2. Problem description

In general, an information system $\mathcal{T}$ with nominal attributes is described as a data table like Table 1. The toy table consists of six objects, i.e. $U = \{o_1, o_2, \ldots, o_6\}$, each of which includes five attributes, $A = \{\text{“}Atrr.1\text{”}, \ldots, \text{“}Attr.4\text{”}, \text{“}Class\text{”}\}$. More specifically, in this example, the attribute set A can be subdivided into four nominal attributes, i.e. $C = \{\text{“}Attr.1\text{”}, \ldots, \text{“}Attr.4\text{”}\}$, and one descision attribute, $D = \{\text{“}Class\text{”}\}$, that is, $A = \{C \cup D\}$.

Given an information system $\mathcal{T}$, the primary goal of feature selection is to find a minimum core subset of features, i.e., C^*, to maintain the capacity which is held by the entire feature set C. Due to the particularity of the nominal data environment, we must first solve the problem of measuring the similarity (distance) of nominal values, for example, how to evaluate the similarity of value “A” and “B”, denoted as δ(“A”,“B”), in Table 1. Also, the second problem is how to minimize the size of the attribute set to keep the capacity.

Roughly speaking, the decision attribute splits the entire universe into space with granular representation U/D, then we need to find a feature set C^* which minimizes the cardinality of C and maximize the approximation quality. Then, the attribute reduction problem solving by Pawlak’s rough set model can be defined as,

$$\begin{aligned} &\max_{C} \min_{|C|} \quad \gamma_C(D) \\ &\text{s.t.} \qquad \delta(v, w) = \begin{cases} 0, & \text{if } v = w, \\ 1, & \text{otherwise} \end{cases}, \quad \forall v, w \in V_a, a = 1, 2, \ldots, n. \end{aligned} \tag{7}$$

where V_a represent the value domain of attribute a, and $\delta(\cdot, \cdot)$ is a distance measure for nominal values, n is the total number of objects. As is discussed in Section 1, the metric $\delta(v, w)$ above would lead to the information loss. Therefore, we transformed the original problem (7) into the following problem, i.e.,

$$\begin{aligned} &\max_{C} \min_{|C|} \quad \gamma_C(D) \\ &\text{s.t.} \quad \delta(v, w) \geq 0, \quad \forall v, w \in V_a, a = 1, 2, \ldots, n. \end{aligned} \tag{8}$$

It has been proved that the problem (7), i.e., computing a minimal reduction of an information system, is an NP-hard problem [41].

4. Modeling

In this section, we first provide a similarity measure for nominal values, and employ the measure to construct the indiscernibility relationship. Then, we propose a ψ-neighborhood rough set model. Finally, we design an attribute reduction algorithm for nominal features based on the ψ-neighborhood rough set model.

4.1. Nominal objects similarity measure

Unlike VDM [42], MVDM [43], we have already considered the importance of the features and objects under the rough set framework so that we will ignore the weights in the original value difference metric definition. In the stage of nominal value similarity calculation, we pay our attention to the statistic difference of nominal values decided by the decision feature.

Definition 6 (Value Frequency Difference Metric, VFDM)**.** Given an information system $\mathcal{T}$, we denote $\mathcal{F}_{ai}^{(j)}$ as the number of times that the value $V_{ai}(V_{ai} \in V_a, i = 1, 2, \ldots, |V_a|)$ in feature $a(a \in A)$ was classified into class j, and denote $\mathcal{F}_{ai}$ as the total number of times V_{ai} occurred in system $\mathcal{T}$. The frequency of value V_{ai} with label j is defined as $\xi_{V_{ai}}^{(j)} = \mathcal{F}_{ai}^{(j)} / \mathcal{F}_{ai}$. Then, the value frequency difference metric for value pair $\{(V_{ak}, V_{al}) | \forall k, l < |V_a|\}$ is,

$$\delta^{(a)}(V_{ak}, V_{al}) = \sum_{j} \left\| \xi_{V_{ak}}^{(j)} - \xi_{V_{al}}^{(j)} \right\|^t, \quad t \in 2^{\{\mathbb{Z}^+ \cup 0\}} \tag{9}$$

The core idea behind this definition is that we wish to connect the values closely which occur with the same relative frequency for all classes.

Definition 7 (Relative Object Dissimilarity Metric, RODM)**.** The dissimilarity of object pair $\{(o_i, o_j) | \forall i, j < n\}$ consists of the difference existing in all of the attributes value pairs. To simplify the notation, we define the auxiliary function of the attribute a as $\tilde{f}_a : 2^U \to V_a, a = 1, 2, \ldots, n,$ where V_a is the value domain of the attribute a. Then, the relative object dissimilarity metric of two object (U_i, U_j) is defined as,

$$\Delta(U_i, U_j) = \sum_{a \in A} \delta^{(a)}\left(\tilde{f}_a(U_i), \tilde{f}_a(U_j)\right) \tag{10}$$

Theorem 1. *Given $\forall \boldsymbol{b}, \boldsymbol{d} \in U$, the relative object dissimilarity metric $\Delta(\boldsymbol{b}, \boldsymbol{d})$ satisfies the following properties:*

1. $\Delta(\boldsymbol{b}, \boldsymbol{d}) \geq 0$
2. $\Delta(\boldsymbol{b}, \boldsymbol{d}) = \Delta(\boldsymbol{d}, \boldsymbol{b})$
3. $\Delta(\boldsymbol{b}, \boldsymbol{b}) = 0$
4. $\Delta(\boldsymbol{b}, \boldsymbol{d}) + \Delta(\boldsymbol{d}, \boldsymbol{c}) \geq \Delta(\boldsymbol{b}, \boldsymbol{c})$

Proof.

(1) For any $t \in 2^{\{\mathbb{Z}^+\}}$, i.e. t is an even number, then the even power of any number is greater than or equal to 0, that is, $|| \cdot ||^t \geq 0$; For $t = 2^{\{0\}} = 1$, we have $| \cdot | \geq 0$. The sum of the non-negative values is non-negative, therefore, $\Delta(\boldsymbol{b}, \boldsymbol{d}) \geq 0$.

(2) Obviously, equation $\delta^{(a)}(\tilde{f}_a(\boldsymbol{b}), \tilde{f}_a(\boldsymbol{d})) = \delta^{(a)}(\tilde{f}_a(\boldsymbol{d}), \tilde{f}_a(\boldsymbol{b}))$ is always true, then, we have $\Delta(\boldsymbol{d}, \boldsymbol{b}) = \Delta(\boldsymbol{b}, \boldsymbol{d})$.

(3) According to the Eq. (9), if $\boldsymbol{d} = \boldsymbol{b}$, then $\delta^{(a)}(\boldsymbol{b}, \boldsymbol{d}) = 0$. Hence, we have the equation $\Delta(\boldsymbol{b}, \boldsymbol{d}) = 0$, when equation $\boldsymbol{d} = \boldsymbol{b}$ is satisfied.

(4) According to the Eq. (10), we have,

$$\Delta(\boldsymbol{b}, \boldsymbol{d}) = \sum_{a} \sum_{j} \left\| \xi_{V_{ak}}^{(j)} - \xi_{V_{al}}^{(j)} \right\|^t$$

$$\Delta(\boldsymbol{d}, \boldsymbol{c}) = \sum_{a} \sum_{j} \left\| \xi_{V_{al}}^{(j)} - \xi_{V_{ap}}^{(j)} \right\|^t$$

$$\Delta(\boldsymbol{b}, \boldsymbol{c}) = \sum_{a} \sum_{j} \left\| \xi_{V_{ak}}^{(j)} - \xi_{V_{ap}}^{(j)} \right\|^t$$

Norms have the following properties: for matrics $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, it satisfies

$$||\boldsymbol{A}||^t + ||\boldsymbol{B}||^t \geq ||\boldsymbol{A} + \boldsymbol{B}||^t$$

If we let $a' = \xi_{V_{ak}}^{(j)} - \xi_{V_{al}}^{(j)}$, $b' = \xi_{V_{al}}^{(j)} - \xi_{V_{ap}}^{(j)}$, and $c' = \xi_{V_{ak}}^{(j)} - \xi_{V_{ap}}^{(j)}$, then we have $||a'|| + ||b'|| \geq ||a' + b'|| = ||c'||$. It means that,

$$\sum_a \sum_j \left(\left\| \xi_{V_{ak}}^{(j)} - \xi_{V_{al}}^{(j)} \right\|^t + \left\| \xi_{V_{al}}^{(j)} - \xi_{V_{ap}}^{(j)} \right\|^t \right) \geq \sum_a \sum_j \left\| \xi_{V_{ak}}^{(j)} - \xi_{V_{ap}}^{(j)} \right\|^t$$

Hence, $\Delta(\boldsymbol{b}, \boldsymbol{d}) + \Delta(\boldsymbol{d}, \boldsymbol{c}) \geq \Delta(\boldsymbol{b}, \boldsymbol{c})$. □

Therefore, RODM can be used as a distance metric with attribute value statistic information, compared with the Hamming distance. It can also be seen that RODM is equivalent to the Hamming distance when the value distance, which is larger than 0, is limited to 1.

Theorem 2. *The Hamming distance measure is a special case of the relative object dissimilarity metric (RODM).*

Proof. If we place a constraint on VFDM, i.e.,

$$\delta^{(*a)}(V_{ak}, V_{al}) = \begin{cases} \mathbb{I}(\delta^{(a)}(V_{ak}, V_{al}) \geq 0), & \text{if } V_{ak} \neq V_{al}; \\ 0, & \text{Otherwise.} \end{cases} \tag{11}$$

where $\mathbb{I}(\cdot)$ is an indicator function with $\mathbb{I}(ture) = 1$ and $\mathbb{I}(false) = 0$. Then, given $\forall \boldsymbol{b}, \boldsymbol{d} \in U$, we have the special form of the RODM $\Delta^*(\ \cdot\ ,\ \ \cdot\ )$ as follows,

$$\begin{aligned} \Delta^*(\boldsymbol{b}, \boldsymbol{d}) &= \sum_{a \in A} \delta^{(*a)}(\tilde{f}_a(\boldsymbol{b}), \tilde{f}_a(\boldsymbol{d})) \\ &= \sum_{a \in A} \mathbb{I}(\delta^{(a)}(\tilde{f}_a(\boldsymbol{b}), \tilde{f}_a(\boldsymbol{d})) \geq 0) \\ &= \sum_{a \in A} \delta^{(a)}(\tilde{f}_a(\boldsymbol{b}), \tilde{f}_a(\boldsymbol{d})) \end{aligned} \tag{12}$$

Obviously, the third line of Eq. (12) is equal to the value frequency difference metric (VFDM), i.e., Eq. (9). □

4.2. ψ-Neighbor rough set model

Unlike Pawlak's rough set model, we employ the object dissimilarity metric (RODM) instead of the Hamming distance metric to measure the dissimilarity of nominal objects. Meanwhile, we propose a quasi-indiscernibility relationship by using this dissimilarity metric to identify the relationship between object pairs.

Definition 8 (Quasi-Indiscernibility Relationships)**.** Given an information system $\mathcal{T}$, a subset P of feature sets A, nominal value dissimilarity metric (VFDM) and object dissimilarity metric (RODM), it induces a quasi-indiscernibility relationship, which is defined as,

$$IND(P_\Delta) \triangleq \{(x, y) \in U \times U | \Delta(x, y) \leq \psi\} \tag{13}$$

Definition 9 (ψ-Neighbor System)**.** Given an information system $\mathcal{T}$ and quasi-indiscernibility relationships on $\mathcal{T}$, it induces a ψ-neighbor system for any object $x \in U$, which is,

$$\psi(x) \triangleq \{y \in U | (x, y) \in IND(P_\Delta)\} \tag{14}$$

Corollary 1. *Given a ψ-neighbor system, it satisfies,*

1. *(non-empty). $\psi(x) \neq \emptyset$, because $x \in \psi(x)$, $\forall x \in U$;*
2. *(symmetry). if $x_i \in \psi(x_j)$ then $x_j \in \psi(x_i)$;*
3. *(closed). $\cup_i \psi(x_i) = U$.*

Definition 10 (ψ-Neighbor Relationships)**.** Given an information system $\mathcal{T}$ and ψ-neighbor system, it induces a cover set $\mathcal{C} = \{\psi(x_i)\}$ of the universe U. Then, the ψ-neighbor relationships is defined as,

$$R_\psi(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \in \psi(x_j); \\ 0, & \text{otherwise;} \end{cases} \tag{15}$$

Definition 11 (Lower and Upper Approximation Operator)**.** Given an information system $\mathcal{T}$, R_ψ is a neighborhood relation defined on U by employing quasi-indiscernibility relationships (Eq. (13)). Then we call $\langle U, R_\psi \rangle$ a ψ-neighbor approximation space. For arbitrary $X \subseteq U$, the lower and upper approximation operator w.r.t. R_ψ is defined as,

$$\begin{aligned} \underline{R_\psi} X &= \{x_i | \psi(x_i) \subseteq X, \forall x_i \in U\} \\ \overline{R_\psi} X &= \{x_i | \psi(x_i) \cap X \neq \emptyset, \forall x_i \in U\} \end{aligned} \tag{16}$$

For arbitrary $X \subseteq U$, it is always true that $\underline{R_\psi}X \subseteq X \subseteq \overline{R_\psi}X$. In general, the sets $\underline{R_\psi}X$ and $U - \overline{R_\psi}X$ are called the positive region and negative region in approximation space $\langle U, R_\psi \rangle$, respectively. The positive region of X shows that the sets which include the equivalence class of each X's element are completely belongs to X. And the negative region of X shows that the sets which include the equivalence class of each X's element is completely not belongs to X.

The gap among the set $\underline{R_\psi}X$ and $\overline{R_\psi}X$ is also called the border region in ψ-neighbor approximation space, i.e.,

$$BR_\psi(X) = \overline{R_\psi}X - \underline{R_\psi}X \tag{17}$$

If $BR_\psi(X) = \emptyset$, then we call the set X is definable in ψ-neighbor approximation space $\langle U, R_\psi \rangle$, otherwise X is a ψ-neighbor rough set.

4.3. Attribute reduction algorithms

Definition 12 (ψ-Neighbor Decision System)**.** Given an information system $\mathcal{T}$, the three triple $\mathcal{T}_\psi = \langle U, R_\psi, D \rangle$ is called ψ-neighbor neighbor decision system. In this system, the universe U is divided into k equivalence classes by the decision attribute D, i.e., $\{X_1, \ldots, X_k\}$. For any $S \subseteq A$, it generates a $R_\psi^{(S)}$ relationship, then, the lower and upper approximation set of D related to S are,

$$\begin{aligned} \underline{R_\psi^{(S)}}D &= \{\underline{R_\psi^{(S)}}X_1, \underline{R_\psi^{(S)}}X_2, \ldots, \underline{R_\psi^{(S)}}X_k\} \\ \overline{R_\psi^{(S)}}D &= \{\overline{R_\psi^{(S)}}X_1, \overline{R_\psi^{(S)}}X_2, \ldots, \overline{R_\psi^{(S)}}X_k\} \end{aligned} \tag{18}$$

where,

$$\begin{aligned} \underline{R_\psi^{(S)}}X &= \{x | \psi_S(x) \subseteq X, \forall x \in U\} \\ \overline{R_\psi^{(S)}}X &= \{x | \psi_S(x) \cap X \neq \emptyset, \forall x \in U\} \end{aligned} \tag{19}$$

It can be seen that Pawlak's rough set model is a special case of the ψ-neighborhood roguh set model. If we let $\psi = 0$, then the quasi-indiscernibility relationship turns into the original definition of indiscernibility relationship in Pawlak's model. In other words, the distance between any pair of objects with nominal value equals to 0, if and only if the objects are identical by applying the relative object dissimilarity metric with $\psi = 0$ to measure object distances. Meanwhile, the quasi-indiscernibility relationships expands the object relationships of binary states, {0, 1}, into a continuous space with similarity interval [0, ψ]. We can control well the degree of similarity between objects, by setting ψ. In summary, through quasi-indiscernibility relationships, we have extended the representation ability of original indiscernibility relationships, that is, it avoids the phenomenon that small changes in the value of the object's nominal attribute affect the large differences in object similarity.

The goal of this paper is to find a subset of the nominal attribute that has similar representation power as decision attributes. To achieve this target, then we define the accuracy of an approximation of U/D w.r.t. S.

Definition 13. Given a ψ-neighbor decision system $\mathcal{T}_\psi = \langle U, R_\psi, D \rangle$, for arbitrary attribute set $S \subseteq A$, the accuracy of an approximation of U/D w.r.t. S is defined as,

$$\gamma_S(D) = |\underline{R_\psi^{(S)}}D| / |U| \tag{20}$$

The accuracy shows the power of the attribute set S to classify each object to its correct label, while the labeled samples are decided by the decision attribute D. Obviously, the range of $\gamma_S(D)$ is between 0 and 1. More specifically, the greater the value of $\gamma_S(D)$, the closer the classification ability of S relative to D.

Definition 14. Given a ψ-neighbor decision system $\mathcal{T}_\psi = \langle U, R_\psi, D \rangle$, $\forall S \subseteq A$ and $\forall a' \in S$, if $\gamma_S(D) > \gamma_{S-a'}(D)$, then we call the attribute a' is necessary for the attribute set S relative to decision attribte D, otherwise not necessary. If the attribute set S satisfies the following condition,

1. $\gamma_S(D) = \gamma_A(D)$;
2. $\forall a' \in S, \quad \gamma_S(D) > \gamma_{S-a'}(D)$

then the attribute set S is called a reduction relative to the attribute set A.

The goal of attribute reduction is to remove the redundant attributes which are not necessary for classifying objects to its true label. The first condition guarantees that the attribute set S have the same distinguishability as the whole attribute set A. And the second condition shows that each attribute a' in a reduction S is necessary.

4.3.1. A heuristic two-stage attribute reduction algorithm

As is discussed in the previous section, it shows that computing a minimal reduction of an information system is an NP-hard problem. Our former works [30,44] show that it is essential to develop some heuristic methods to deal with it. In this paper, we construct a heuristic two-stage attribute reduction algorithm for nominal data, in which splits the original problem into two sub-procedures.

In the first stage, by employing mutual information metrics, we use a heuristic method to search the best attribute path, which generates a candidate pool with attribute priorities as the inputs of the follow-up procedure. In general, the mutual information of two random variables is a measure of the mutual dependence between the two variables. Given a ψ-neighbor decision system $\mathcal{T}_\psi = \langle U, R_\psi, D\rangle, \forall S \subseteq C,$ and $\forall a' \in C - S,$ the increment of the mutual information is,

$$I(S \cup \{a'\}, D) - I(S, D) = H(D|S) - H(D|S \cup \{a'\})$$

Here, random variable S represents the objects that only consists of attribute set S. And the random variable D represents the decision values. $I(\,\cdot\,,\,\cdot\,)$ is the mutual information and $H(\,\cdot\,|\,\cdot\,)$ is conditional entropy. According to information theory, the more significant the increment, the higher the importance of a' to D given the condition of S. Thus, we have the following definition.

Definition 15. Given a ψ-neighbor decision system $\mathcal{T}_\psi = \langle U, R_\psi, D\rangle, \forall S \subseteq C,$ and $\forall a' \in C - S,$ the priority of attribute a' is defined as,

$$AP(a', S, D) = H(D|S) - H(D|S \cup \{a'\}) \tag{21}$$

where $H(\,\cdot\,|\,\cdot\,)$ is a conditional entropy. With the ψ-neighbor relationship R_ψ, the calculation process of $H(\,\cdot\,|\,\cdot\,)$ is showed as follows. For example, we calculate $H(D|S)$ like this,

$$\begin{aligned} H(D|S) &= -\sum_{o_s \in \{U^S\}} p(S = o_s) H(D|S = o_s) \\ &= -\sum_{o_s \in \{U^S\}} p(S = o_s) \sum_{d \in D} p(D = d|S = o_s) \log p(D = d|S = o_s) \end{aligned} \tag{22}$$

where $\{U^S\}$ is the set of objects with attribute set S; S, D is a random variable for the attribute set and the decision attribute, respectively.

In the second stage, according to the second condition in Definition 14, we will calculate the necessary of the attribute from the priority ordered pool which is generated in the first stage. To facilitate the calculation, we define the importance of the attribute a' as,

$$SIG(a', S, D) = \gamma_{S \cup \{a'\}}(D) - \gamma_S(D) \tag{23}$$

Then, we will choose the maximum necessary attribute to add to the reduction. Finally, if the first condition in Definition 14 is satisfied in the reduction, we will stop the attribute reduction algorithm, otherwise, go to the first stage.

A detailed description of the heuristic two-stage attribute reduction algorithm (HTSAR) is given in Algorithm 1.

5. Experiments

The empirical study of the proposed HTSAR is given in this section. Specifically, we first set up the experiments by preparing the experimental datasets and then evaluate the performance concerning the quality of the algorithm and its novelty.

5.1. Data preparation and experiment environment

We evaluate our method on six nominal datasets: Weather, Zoo, Soybean, Dermatology, Lymphography, and Breast-Caner(abbr. Breast), which are collected from UCI machine learning repository.[1] The attribute reduction algorithm HTSAR and its competitors are implemented in MATLAB, where the computationally large part of the algorithm is implemented by using C++ to achieve and finally embedded into MATLAB. All the experiments are executed at a PC with CPU 3.4GHz and 14GB memory.

5.2. The representation ability at different scale

Firstly, we employ the RODM measure to calculate the distribution of object similarities for the six UCI datasets. The RODM method maps the distance of nominal objects from a discrete space to a continue space. Fig. 1 shows the normalized distribution of objects in the six UCI datasets, and Fig. 2 shows the similarity distribution of objects for all datasets, respectively. As can be seen from Fig. 12, the distribution of data is complicated. If we use the equivalence relation based on the

[1] http://archive.ics.uci.edu/ml.

Algorithm 1 A Heuristic Two-Stage Attribute Reduction Algorithm.

```
Input: $\mathcal{T} = \langle U, C \cup D, V, f \rangle$, neighbor size: $\psi$, pool size: $\kappa$, tolerance: $\epsilon$.
Output: reduction $S$.
//check the value of $\psi$
if ($\underline{R_{\psi}^{(C)}}D \neq |U|$) then
    return “the value of $\psi$ is wrong.”;
end if
$S \leftarrow \emptyset$
repeat
    //first stage
    Initialize the attribute pool $AP[1.\kappa]$;
    for $a \in \{C - S\}$ do
        $AP'[a] \leftarrow$ Eq. (21);
    end for
    sort $AP'[1.|C - S|]$ with descending order;
    $AP[1.\kappa] \leftarrow AP'[1.\kappa]$;
    select candidate attributes $\alpha[1.\kappa]$ with $AP[1.\kappa]$
    //second stage
    for $i = 1.\kappa$ do
        calculate $SIG[\alpha[i]] \leftarrow$ Eq. (23)
    end for
    sort $SIG[1.\kappa]$;
    select the attribute $a'$ with maximum value $SIG[a']$;
    update reduction $S \leftarrow S \cup \{a'\}$;
until (1− Eq. (20)$\leq \epsilon$)
return $S$.
```

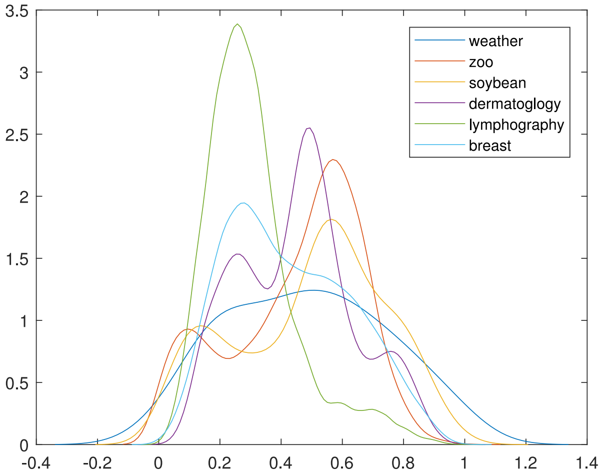


Fig. 1. Normalized object distance distribution for the six UCI datasets.

Hamming metric to construct the rough set model, then the model is difficult to reflect the intrinsic characteristics of the nominal data, because the Hamming measure calculated distance is discrete, which will lead to information loss. Therefore, the existing rough set models for nominal data cannot represent data well.

Compared to Pawlak’s rough set model, the ψ-neighborhood rough set model can represent data at different scales. In other words, the model can choose an optimal value of the parameter ψ to maintain the characteristics of the original data. By adjusting the value of ψ, the indiscernibility relationship of any two objects will be changed. Therefore, the parameter ψ plays an important role in the ψ-neighbor decision system, and controls the representation ability of the rough set model for nominal data via the lower and the upper approximation operator. If we set the value of ψ to a large enough number in the HTSAR algorithm, then all objects become a set that each member is equivalent to another one. As a consequence, the algorithm cannot get a selected subset of attributes, because the relations between all object pairs are indistinguishable. That is why we first check the validity of the value of ψ, because it is a meaningless action that searches a smallest set of attributes in a system which lacks the data-representability.

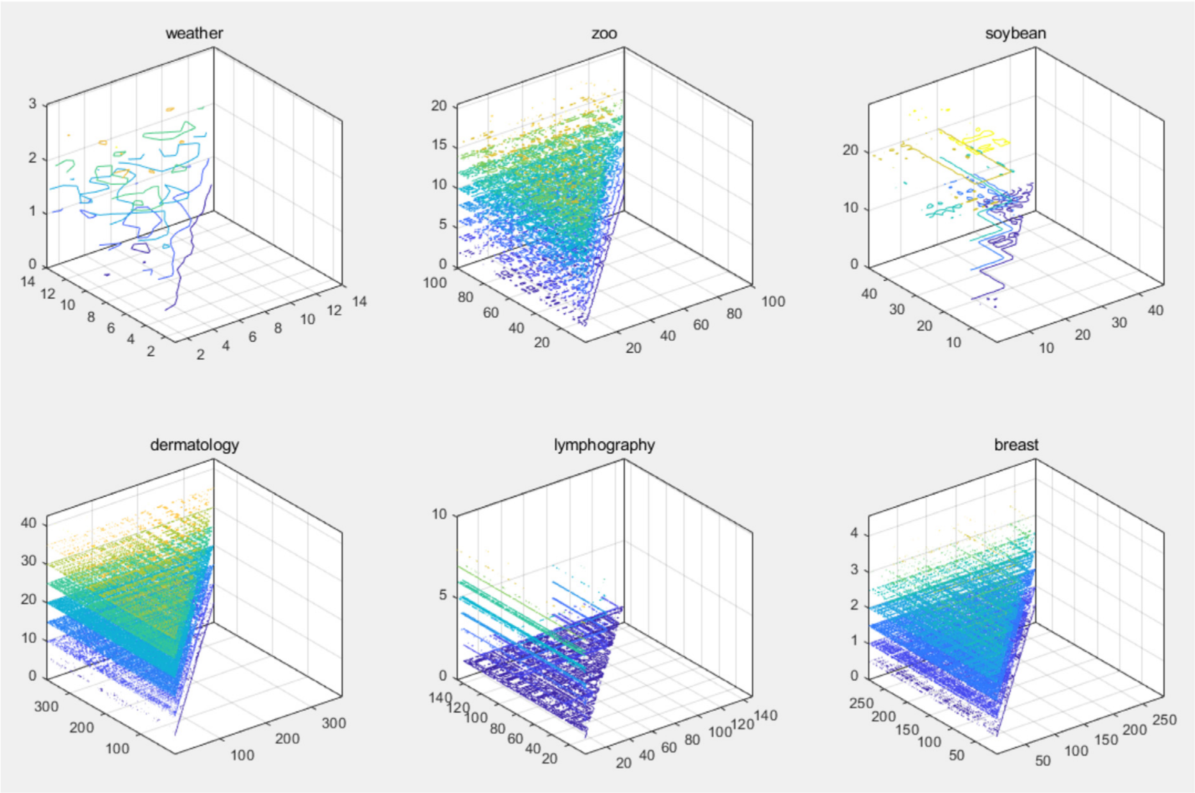


Fig. 2. The 3D contours of object pair similarities. The *x*, *y*-axis label represents the object from the data set named by the sub figure title. The *z*-axis label is the distance between pairs of data objects.

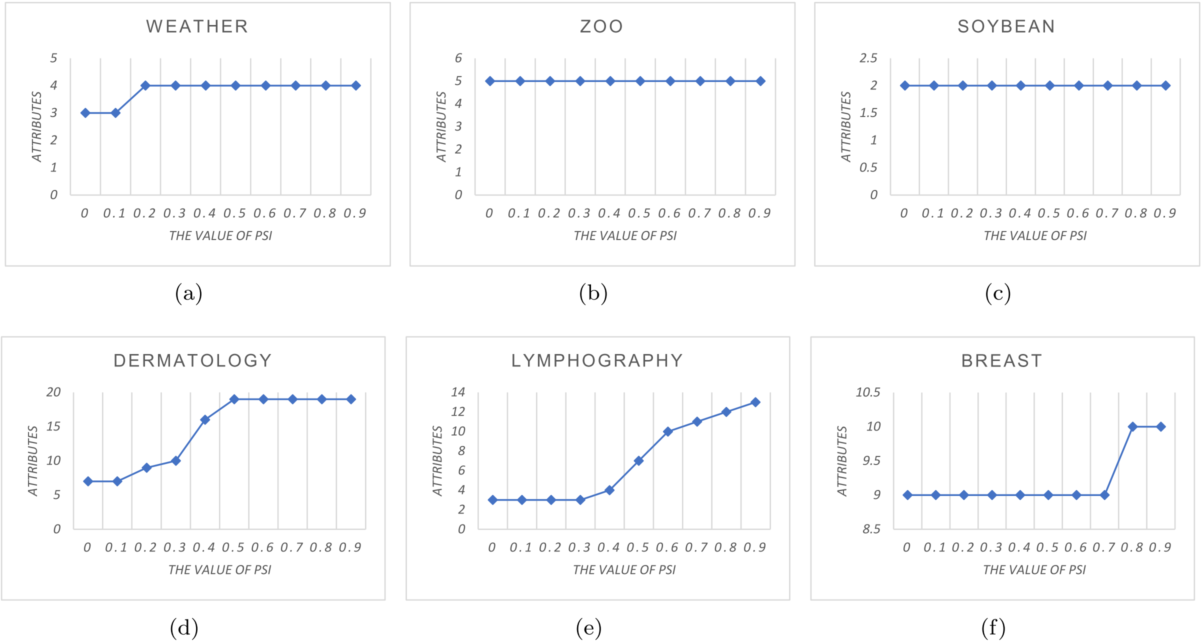
In general, we hope the optimal value of the parameter ψ is in line with the data inherent requirements. Fig. 3 displays the cardinality variation of the selected attribute subset accompanying with the changed ψ. When the parameter $\psi = 0$ is determined at the x-axis in each subgraph, the ψ-neighborhood rough set model is consistent with the Pawlak rough set model with the Hamming distance. Obviously, as the value of ψ increases, the number of objects which satisfies the indiscernibility relationship also increases. In order to distinguish objects, we need more attributes (features). Therefore, the cardinality of the selected attribute set also increases. This feature selection mechanism is consistent with the human being's recognition. There is a positive correlation relationship between the cardinality of the selected attribute subset and the value of the parameter ψ, until the ψ value affects the ability of the model to distinguish the indiscernibility relationship in objects. From Fig. 3, it can be seen that the ψ value has not yet reached the limit of data representation ability in the dataset Zoo and Soybean, which is significantly different from the traditional rough set model. These results show that the Pawlak rough set model is a special case of the ψ-neighborhood rough set model for nominal data.

The most important thing is that the ψ-neighborhood rough set model constructed a multi-scale representation of the data. It is the most significant difference between the ψ-neighborhood rough set model and other models such as the classic rough set model, the probabilistic rough set model, the fuzzy-rough set model, etc. Fig. 3 also shows the robustness and stability of the ψ-neighborhood rough set model. In extreme circumstances, we randomly add noise to the excluded attribute value, the selected attribute set obtained by ψ-neighborhood rough set model does not be affected, while others rough set model would be affected severely.

5.3. The performance of the HTSAR

Obviously, this kind of multi-scale representation of the ψ-neighborhood rough set model cannot be represented by the existing rough set model. To compare with the attribute reduction algorithm of single-granularity representation rough set model, we set ψ to 0 and select three algorithms, i.e., Pawlak's attribute reduction algorithm (Pawlak) [5], quick attribute reduction algorithm(QAR) [45], and entropy-based attribute reduction(EBAR) [46].

Table 2 demonstrates the feature selection result of HTSAR algorithm compared with Pawlak, QAR, EBAR in the original feature space. From Table 2 we could see clearly that the HTSAR algorithm can get fewer attributes than Pawlak, QAR, and EBAR, which keeps the classification ability unchanged. The following is the specific situation of attribute reduction on each data set. The number of the attributes that HTSAR algorithm selected is (75%, 50%, 40%, 46.67%, 33.3%, 100%), (75.00%,



(a) (b) (c)

(d) (e) (f)

Fig. 3. The relationship between the cardinality of the selected attribute subset and the value of ψ. The variable ψ controls the degree of indiscernibility and also restricts the degree of similarity between objects from the statistic perspective. The optimum value of ψ is related to the data set.

Table 2
The result comparison of attribute reduction.

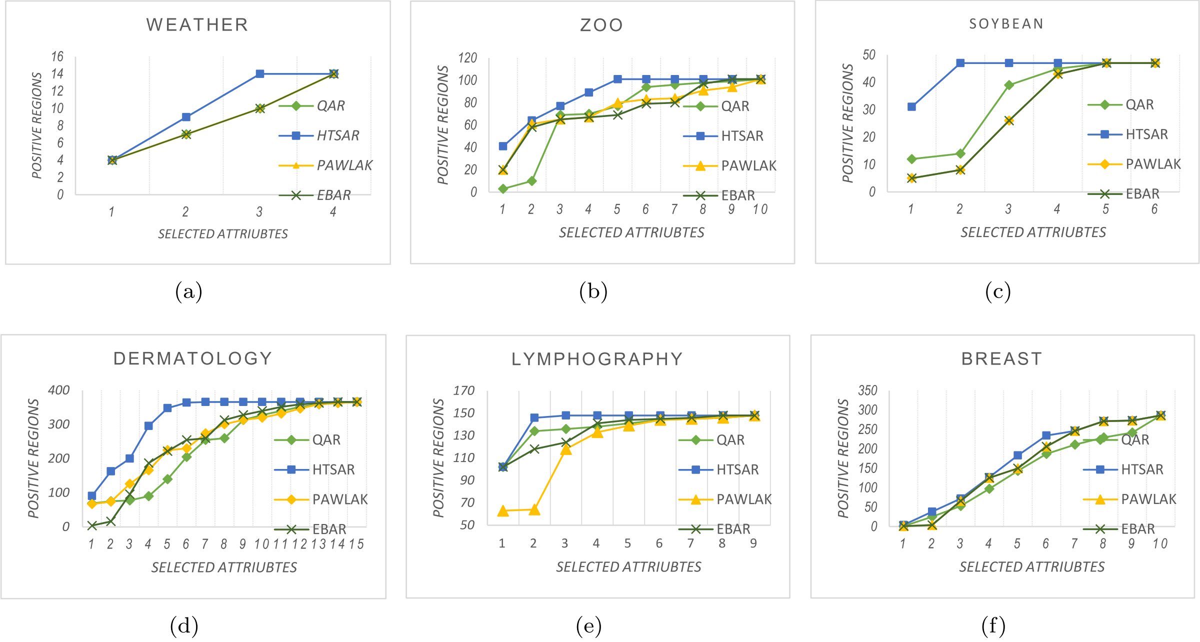
Data Sets	Classes	Features	Number of selected attributes			
			Pawlak	QAR	EBAR	HTSAR
weather	2	4	4	4	4	3
zoo	7	16	10	10	10	5
soybean	4	35	5	4	6	2
dermatology	6	34	15	14	14	7
lymphography	2	18	9	7	6	3
breast	2	9	9	9	9	9

50.00%, 50.00%, 50.00%, 42.86%, 100.00%), (75.00%, 50.00%, 33.33%, 50.00%, 50.00%, 100.00%) for algorithm Pawlak, QAR and EBAR, respectively. In particular, the algorithm HTSAR does not compress the number of attributes for the dataset Breast, which is consistent with the others attribute reduction algorithms.

Fig. 4 shows the positive region changes of the selected attribute set relative to decision attribute after each selected attribute is added. From the Fig. 4 above, we also conclude that HTSAR has the best positive regions, after each selected attribute is added, compared with Pawlak, QAR, EBAR. One of the main reasons is that heuristic information plays an essential role in the choice of attributes in the pool of candidate attributes during the first phase of the HTSAR algorithm. The most significant difference between the HTSAR algorithm and Pawlak, QAR, EBAR, is that each attribute selected by HTSAR is the one with the highest classification accuracy improvement in the candidate pool, while the other methods only select attributes that can elevate classification accuracy, which is not necessarily an optimal choice.

To enhance the efficiency and stability of the algorithm HTSAR, we limit the pool size to 3 in this experiment. Theoretically, if computing power is not a bottleneck, we can dynamically adjust the pool size to the total number of remaining attributes in each iteration. The role of the pool is showed in situations that the mutual information between attributes is inconsistent with the improvement of the positive region. It is another reason that HTSAR can get better performance than the others attribute reduction algorithms. It can be seen from the Fig. 4 that the positive region changes of Pawlak, QAR, EBAR do not have significant regularity because this crucial factor does not be taken into account.

Roughly speaking, the two critical factors of HTSAR are heuristic information and candidate pooling mechanism. Therefore, lacking heuristic information, the attribute reduction algorithm for the selection of candidate attributes is equivalent to a random choice. And lacking the pooling mechanism, the promotion of positive region changes is not optimum for attribute reduction algorithm.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. The positive region changes after each attribute is selected by HTSAR for each iteration.

Table 3
The classification ability of selected attribute sets.

Data sets	*k*nnRODM classification accuracy					
	Features	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
weather	original	0.5000	0.5000	0.6429	0.5714	0.5714
	selected	**0.7143**	**0.7143**	**0.7143**	**0.7143**	**0.6429**
zoo	original	0.9505	0.9505	0.9208	0.9505	0.9406
	selected	**0.9703**	**0.9703**	**0.9505**	**0.9505**	**0.9307**
soybean	original	1.0000	1.0000	1.0000	1.0000	1.0000
	selected	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
dermatology	original	0.9836	0.9836	0.9754	0.9836	0.9809
	selected	**0.9806**	**0.9806**	**0.9822**	**0.9831**	**0.9842**
lymphography	original	0.9865	0.9865	0.9797	0.9797	0.9797
	selected	**0.9932**	**0.9932**	**0.9797**	**0.9797**	**0.9797**
breast	original	0.6748	0.6748	0.6469	0.6783	0.6818
	selected	**0.6748**	**0.6748**	**0.6469**	**0.6783**	**0.6818**

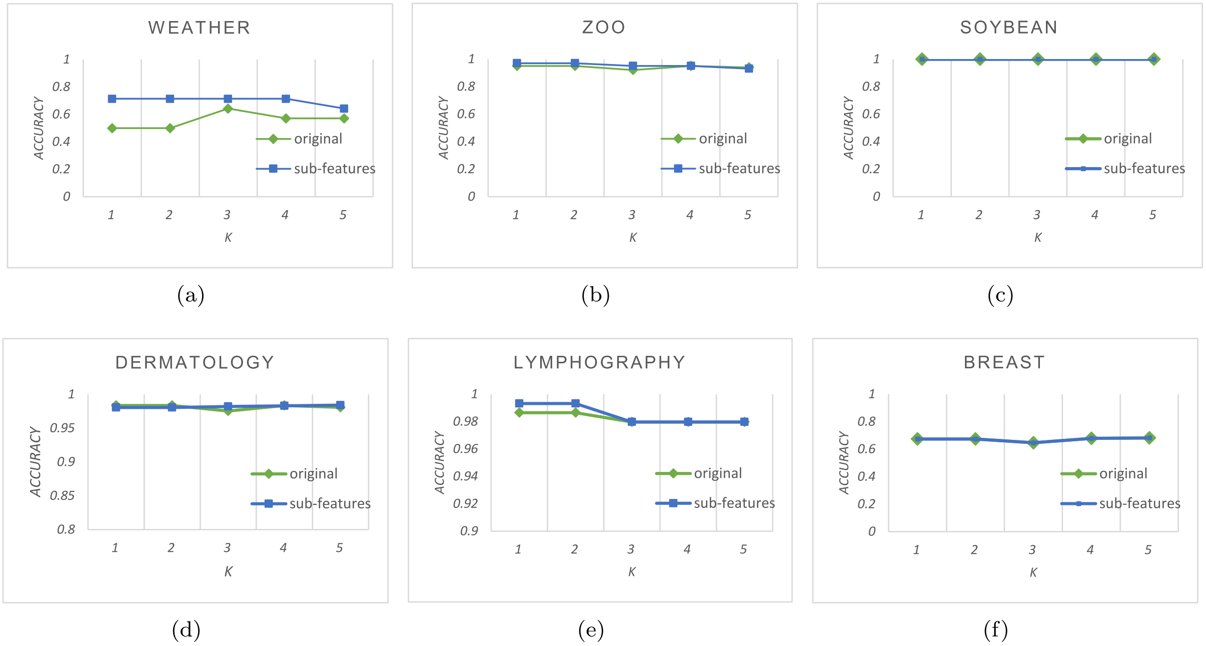
5.4. The quality of the selected attribute set

To test the classification ability of the attribute set obtained by the attribute reduction algorithm HTSAR, we replace the distance measure in *k*NN classification algorithm with RODM, named *k*nnRODM, and then use the 10-fold cross-validation strategy to test the classification ability of selected attribute set. The detailed results are shown in the following Table 3. Since the dataset Breast does not have a minimal reduction of the attribute set, the classification accuracy remains constant. For the others datasets, it is clear that the selected attribute set can not only maintaining the classification ability of the original attribute set but also improving the classification ability, especially the first dataset, i.e., Weather, is more significant. The reason for this increase in accuracy is that HTSAR rejects attribute sets that are equally weighted but not necessary for classification tasks, in other words, the attributes of the rejection interfere with the accuracy of the classification.

Fig. 5 describes the relationship between classification accuracy and the *k* value in *k*nnRODM classification algorithm for whole attribute set and selected attribute set respectively. As can be seen from the figure, the choice of *k* value has a significant effect on the classification accuracy. Also, the optimum choice of *k* value depends on a specific dataset.

To test the performance of the knnRODM, we also use the following methods for the experiments.

- *kNN+Hamming*. This *k*NN model simply classify objects with categorical attributes by using the Hamming distance.
- *C4.5*[47]. This model builds a decision tree to classify the categorical objects, according to the information gain rate.



(a) (b) (c)

(d) (e) (f)

Fig. 5. The relationships between classification accuracy and the value of *k*.

Table 4
The accuracy result of the models performed on the six data sets.

Methods	Accuracy					
	Weather	Zoo	Soybean	Dermatology	Lymphography	Breast
*k*NN+Hamming	0.3571 ± 0.1195	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.4144 ± 0.1456	0.4959 ± 0.0193
C4.5	0.5000 ± 0.3581	0.8955 ± 0.0189	0.9710 ± 0.0106	0.9316 ± 0.0176	0.6574 ± 0.0236	0.7552 ± 0.1870
NB	0.5714 ± 0.2417	0.9089 ± 0.0105	1.0000 ± 0.0000	0.9658 ± 0.0054	0.9196 ± 0.0106	0.7202 ± 0.0021
*k*nnRODM+Reduct	0.5571 ± 0.0598	0.9426 ± 0.0129	1.0000 ± 0.0000	0.9814 ± 0.0036	0.9824 ± 0.0037	0.6713 ± 0.0140

- *Naive Bayes(NB)*[48]. This model assumes that the attributes are independent and need training data to fit the parameters of the model. The trained model will directly output the class label for the new object without comparing it with the neighbors.

Table 4 demonstrates the results of the classification accuracy over the six UCI datasets, respectively. All results are performed under the settings with the 10-fold cross validation strategy and the parameter k ranges from 1 to 5. And the average performances were reported with the style of $\mu \pm \sigma^2$, where μ represents the average and σ^2 is the deviation. The table shows that the classifier knnRODM with reduct attributes completely outperform the classifier kNN+Hamming over the six UCI data sets. The table also shows that the performance of the classifier knnRODM is better than C4.5 and NB, except for the dataset Breast. In this experiment, the loss of information caused by the Hamming distance leads to a decrease in the performance of the classifier. At the worst case, the classification accuracy of kNN+Hamming is 0.0000, which fully demonstrates the unreliability of Hamming distance measure for classification tasks. Therefore, the loss of information is the main source of defects for the Hamming distance based Rough set model. Moreover, it can be seen that reduction attributes have the same classification ability as full attributes. In summary, the experimental results show the reliability of our proposed RODM and the effectiveness of the attribute reduction algorithm, HTSAR.

6. Conclusion

How to effectively represent data in a friendly form is a difficult task due to the complexity of data. This work proposed a ψ-neighborhood rough set model to map data to a multi-scale representation by using the RODM metric for nominal data. The ψ-neighborhood rough set model extends the classic rough set model to a more robust, representative, and effective model which is close to the characteristic of nominal data. Besides, the strategy of dynamically adjusting the indiscernibility relationship of data based on the parameter ψ value provides a new viewpoint to detect latent knowledge from nominal data and extends the rough set model from the data-driven perspective. Base on the ψ-neighbor rough set model, we proposed a heuristic two-stage attribute reduction algorithm(HTSAR) to perform attribute reduction tasks. Experiments show that the ψ-neighborhood rough set model can take advantage of more potential knowledge in nominal data and achieve better performance for attribute reduction than the existing rough set model.

Several aspects of the new method are worth investigating in further depth, including the strategy of searching an optimal value of ψ according to a specific dataset, the dissimilarity measure for nominal data, and the discovery of more complex relationships in the nominal data, etc.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Sheng Luo: Writing - original draft, Writing - review & editing, Conceptualization, Formal analysis, Methodology, Data curation, Validation. **Duoqian Miao:** Conceptualization, Formal analysis, Supervision, Writing - review & editing, Funding acquisition. **Zhifei Zhang:** Writing - review & editing, Data curation, Investigation. **Yuanjian Zhang:** Writing - review & editing, Validation. **Shengdan Hu:** Writing - review & editing, Validation.

Acknowledgements

The authors thank both the editors and the anonymous referees for their valuable suggestions, which substantially improved this paper. This work is supported by National Key R&D Program of China (Grant No. 213), the National Science Foundation of China (Grant No. 61673301, 61502259), and Major Project of Ministry of Public Security (Grant No. 20170004).

References

[1] T. Sapatinas, Discriminant analysis and statistical pattern recognition, J. R. Stat. Soc. 35 (3) (2005) 324–326, doi:10.1111/j.1467-985x.2005.00368_10.x.

[2] K. Pearson, On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dubl. Philos. Mag. J. Sci. 2 (11) (1901) 559–572.

[3] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (3) (1994) 287–314 Higher Order Statistics, doi:10.1016/0165-1684(94)90029-9.

[4] Y. Bengio, Learning deep architectures for AI, Now Found. Trends, 2009, doi:10.1561/2200000006.

[5] Z. Pawlak, Rough set theory and its applications, J. Telecommun. Inf. Technol. 3 (3) (1998) 7–10.

[6] H.H. Bock, E. Diday (Eds.), Analysis of Symbolic Data, Springer Berlin Heidelberg, 2000, doi:10.1007/978-3-642-57155-8.

[7] S. Wong, W. Ziarko, Comparison of the probabilistic approximate classification and the fuzzy set model, Fuzzy Sets Syst. 21 (3) (1987) 357–362, doi:10.1016/0165-0114(87)90135-7.

[8] W. Ziarko, Probabilistic approach to rough sets, Int. J. Approximate Reasoning 49 (2) (2008) 272–284 Special Section on Probabilistic Rough Sets and Special Section on PGMâ06, doi:10.1016/j.ijar.2007.06.014.

[9] Y. Yao, S. Wong, A decision theoretic framework for approximating concepts, Int. J. Man Mach. Stud. 37 (6) (1992) 793–809, doi:10.1016/0020-7373(92)90069-W.

[10] D. Slezak, W. Ziarko, The investigation of the Bayesian rough set model, Int. J. Approx. Reason. 40 (1) (2005) 81–91 Data Mining and Granular Computing, doi:10.1016/j.ijar.2004.11.004.

[11] Y. Yao, Probabilistic rough set approximations, Int. J. Approx. Reason. 49 (2) (2008) 255–271 Special Section on Probabilistic Rough Sets and Special Section on PGMâ06, doi:10.1016/j.ijar.2007.05.019.

[12] T. Beaubouef, F.E. Petry, G. Arora, Information-theoretic measures of uncertainty for rough sets and rough relational databases, Inf. Sci. 109 (1) (1998) 185–195, doi:10.1016/S0020-0255(98)00019-X.

[13] D. Miao, L. Hou, A comparison of rough set methods and representative inductive learning algorithms, Fundam. Inform. 59 (2–3) (2004) 203–219.

[14] N. Zhong, A. Skowron, A rough set-based knowledge discovery process, Int. J. Appl. Math. Comput. Sci. 11 (2001) 603–619.

[15] M. Kryszkiewicz, Rough set approach to incomplete information systems, Inf. Sci. 112 (1) (1998) 39–49, doi:10.1016/S0020-0255(98)10019-1.

[16] Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inf Sci (Ny) 111 (1) (1998) 239–259, doi:10.1016/S0020-0255(98)10006-3.

[17] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Trans. Knowl. Data Eng. 12 (2) (2000) 331–336, doi:10.1109/69.842271.

[18] S. Greco, B. Matarazzo, R. Slowinski, Rough sets theory for multicriteria decision analysis, Eur. J. Oper. Res. 129 (1) (2001) 1–47, doi:10.1016/S0377-2217(00)00167-3.

[19] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int J. Gen. Syst. 17 (2–3) (1990) 191–209, doi:10.1080/03081079008935107.

[20] S. Pei, Q. Hu, Partially monotonic decision trees, Inf. Sci. 424 (2018) 104–117, doi:10.1016/j.ins.2017.10.006.

[21] Y. Leung, M.M. Fischer, W.-Z. Wu, J.-S. Mi, A rough set approach for the discovery of classification rules in interval-valued information systems, Int. J. Approx. Reason. 47 (2) (2008) 233–246, doi:10.1016/j.ijar.2007.05.001.

[22] B. Sun, Z. Gong, D. Chen, Fuzzy rough set theory for the interval-valued fuzzy information systems, Inf. Sci. 178 (13) (2008) 2794–2815, doi:10.1016/j.ins.2008.03.001.

[23] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594, doi:10.1016/j.ins.2008.05.024.

[24] Q. Hu, D. Yu, W. Pedrycz, D. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans Knowl Data Eng 23 (11) (2011) 1649–1667, doi:10.1109/tkde.2010.260.

[25] W. Wu, W. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Inf. Sci. 159 (3) (2004) 233–254, doi:10.1016/j.ins.2003.08.005.

[26] G.Y. Wang, J. Zhao, J. An, Y. Wu, A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundam. Inform. 68 (3) (2005) 289–301.

[27] W. Zhu, F.Y. Wang, On three types of covering-based rough sets, IEEE Trans. Knowl. Data Eng. 19 (8) (2007) 1131–1144, doi:10.1109/TKDE.2007.1044.

[28] Y. Yao, Y. Zhao, J. Wang, On reduct construction algorithms, Trans. Comput. Sci. 4062 (2008) 100–117, doi:10.1007/978-3-540-87563-5_6.

[29] J. Mi, Y. Leung, W. Wu, An uncertainty measure in partition-based fuzzy rough sets, Int. J. Gen. Syst. 34 (1) (2005) 77–90, doi:10.1080/03081070512331318329.

[30] D. Miao, G. Hu, A heuristic algorithm for reduction of knowledge, Journal of Computer Research & Development 36 (6) (1999) 681–684. (In Chinese).

[31] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: Intelligent Decision Support, Springer, 1992, pp. 331–362.

[32] S. Li, T. Li, D. Liu, Dynamic maintenance of approximations in dominance-based rough set approach under the variation of the object set, Int. J. Intell. Syst. 28 (8) (2013) 729–751.

[33] X. Jia, L. Shang, B. Zhou, Y. Yao, Generalized attribute reduct in rough set theory, Knowl. Based Syst. 91 (2016) 204–218.

[34] A. Ahmad, L. Dey, A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set, Pattern Recognit. Lett. 28 (1) (2007) 110–118, doi:10.1016/j.patrec.2006.06.006.

[35] H. Jia, Y. Cheung, J. Liu, A new distance metric for unsupervised learning of categorical data, IEEE Trans. Neural Netw. Learn. Syst. 27 (5) (2016) 1065–1079, doi:10.1109/TNNLS.2015.2436432.

[36] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, Y. Ou, Coupled nominal similarity in unsupervised learning, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, in: CIKM '11, ACM, New York, NY, USA, 2011, pp. 973–978, doi:10.1145/2063576.2063715.

[37] E. Zdravevski, P. Lameski, A. Kulakov, S. Kalajdziski, Transformation of nominal features into numeric in supervised multi-class problems based on the weight of evidence parameter, in: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), 2015, pp. 169–179, doi:10.15439/2015F90.

[38] S. Jian, L. Cao, G. Pang, K. Lu, H. Gao, Embedding-based representation of categorical data by hierarchical value coupling learning, in: International Joint Conference on Artificial Intelligence, 2017, pp. 1937–1943.

[39] S. Peng, Q. Hu, Y. Chen, J. Dang, Improved support vector machine algorithm for heterogeneous data, Pattern Recognit. 48 (6) (2015) 2072–2083, doi:10.1016/j.patcog.2014.12.015.

[40] P. Zhu, Q. Hu, W. Zuo, M. Yang, Multi-granularity distance metric learning via neighborhood granule margin maximization, Inf. Sci. 282 (2014) 321–331, doi:10.1016/j.ins.2014.06.017.

[41] S.K.M. Wong, W. Ziarko, On optimal decision rules in decision tables, Bull. Polish Acad. Sci.Math. 33 (11–12) (1985) 693–696.

[42] C. Stanfill, D. Waltz, Toward memory-based reasoning., Commun. ACM 29 (12) (1986) 1213–1228, doi:10.1145/7902.7906.

[43] S. Cost, S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features, Mach. Learn. 10 (1) (1993) 57–78, doi:10.1007/bf00993481.

[44] X. Yue, Y. Chen, D. Miao, J. Qian, Tri-partition neighborhood covering reduction for robust classification, Int. J. Approx. Reason. 83 (2017) 371–384, doi:10.1016/j.ijar.2016.11.010.

[45] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: areview, Appl. Soft. Comput. 9 (1) (2009) 1–12, doi:10.1016/j.asoc.2008.05.006.

[46] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. Data Eng. 16 (12) (2004) 1457–1471, doi:10.1109/tkde.2004.96.

[47] J.R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[48] D.J. Hand, K. Yu, Idiot's Bayes – not so stupid after all? Int. Stat. Rev. 69 (3) (2001) 385–398.