Fast and Robust Attribute Reduction Based on the Separability in Fuzzy Decision Systems

Meng Hu, Eric C. C. Tsang, Yanting Guo, and Weihua Xu

***Abstract*—Attribute reduction is one of the most important preprocessing steps in machine learning and data mining. As a key step of attribute reduction, attribute evaluation directly affects classification performance, search time, and stopping criterion. The existing evaluation functions are greatly dependent on the relationship between objects, which makes its computational time and space more costly. To solve this problem, we propose a novel separability-based evaluation function and reduction method by using the relationship between objects and decision categories directly. The degree of aggregation (DA) of intraclass objects and the degree of dispersion (DD) of between-class objects are first defined to measure the significance of an attribute subset. Then, the separability of attribute subsets is defined by DA and DD in fuzzy decision systems, and we design a sequentially forward selection based on the separability (SFSS) algorithm to select attributes. Furthermore, a postpruning strategy is introduced to prevent overfitting and determine a termination parameter. Finally, the SFSS algorithm is compared with some typical reduction algorithms using some public datasets from UCI and ELVIRA Biomedical repositories. The interpretability of SFSS is directly presented by the performance on MNIST handwritten digits. The experimental comparisons show that SFSS is fast and robust, which has higher classification accuracy and compression ratio, with extremely low computational time.**

***Index Terms*—Attribute reduction, fuzzy decision systems, fuzzy membership, separability.**

I. Introduction

ATTRIBUTE reduction aims to reduce irrelevant, redundant, and inconsistent attributes from the original attribute set so as to better mine the potential rules between knowledge and tasks, and help decision making and prediction. It has been widely used in machine learning, data mining, and pattern recognition. An outstanding technology of attribute selection cannot only reduce the cost of classification and regression tasks but also improve the performance of the tasks and reduce storage space [2], [14].

In practical applications, many data with high dimensions and high noise are fuzzy and uncertain. Many attribute reduction methods in fuzzy data were proposed, such as fuzzy rough sets (FRSs) [10], [16], [24], [25]; kernel machine learning [5], [10], [25]; information entropy [3], [6]; and neighborhood relation [1]–[4], [6]. Fuzzy sets were first proposed by Zadeh in 1965 [13], which have been applied to neural networks [18], [19]; support vector machines [20], [21]; logic controllers [23]; ensemble learning [35], [36]; and attribute selection [16], [22]. By combining the fineness of the boundary description of fuzzy sets with the objectivity of knowledge expression of rough sets [15], [37], [38], FRSs describe the certain and possible membership degrees of objects with respect to categories by two approximate operators.

Dubois and Prade [26], [27] first introduced the fuzzy sets into the upper and lower approximations of rough sets to propose FRS. Wu *et al.* [28] extended the fuzzy equivalence relation to the general binary fuzzy relation and formed more generalized FRSs. Some researchers further studied attribute reduction based on the FRS [16], [22], [24], [25]. Hu *et al.* [10] proposed two types of kernelized FRS by integrating kernel functions with FRSs. Chen *et al.* [25] combined the Gaussian kernel function with FRSs to propose parameterized attribute reduction. In [5], the fuzzy kernel assignment of the combined kernel and the ideal kernel was minimized, and then the attribute with high assignment value was selected, whereas it would be eliminated. To calculate the distinguishing ability of similarity relations, Yager [31] provided an extension of the Shannon entropy based on the fuzzy similarity relation to measure the significance of attribute subsets. Hernĺćndez and Recasens [32] presented the joint entropy and conditional entropy based on the Yager entropy to learn fuzzy decision trees. The entropy-based attribute reduction has been widely used in gene data expression and medical diagnosis [29], [30], [46]. The attribute reduction of joint mutual information and FRS was studied in literature [29], which has been applied in cancer classification. In 2011, Hu *et al.* [3] constructed a neighborhood mutual information measure by combining the Shannon information entropy with the neighborhood relation, then used it to evaluate the significance of continuous and discrete features. In 2016, Wang *et al.* [33] established a fuzzy neighborhood rough set model and defined the dependency between fuzzy decision and conditional attributes to measure the significance of attributes

Manuscript received January 6, 2020; revised July 28, 2020 and October 29, 2020; accepted November 23, 2020. This work was supported in part by the Macau Science and Technology Development Funds under Grant 0019/2019/A1 and Grant 0075/2019/A2; and in part by the Natural Science Foundation of China under Grant 61976245 and Grant 61772002. This article was recommended by Associate Editor C.-F. Juang. *(Corresponding author: Eric C. C. Tsang.)*

Meng Hu, Eric C. C. Tsang, and Yanting Guo are with the Faculty of Information Technology, Macau University of Science and Technology, Macau, China (e-mail: humeng24@sina.com; cctsang@must.edu.mo; ytguosx@sina.com).

Weihua Xu is with the College of Artificial Intelligence, Southwest University, Chongqing 400715, China (e-mail: datongxuweihua@126.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2020.3040803.

Digital Object Identifier 10.1109/TCYB.2020.3040803

and reduce unnecessary attributes. In 2018, the neighborhood discernibility measure [34] was defined by considering consistent samples and discriminate samples with different decisions, which were used to characterize the classification ability of attribute subsets. Neighborhood mutual information [1] was defined by using the cardinality of neighborhood relation to measure the importance of a candidate attribute subset.

The evaluation functions of the above reduction methods all consider the fuzzy relation of objects. The FRS-based attribute reduction [17] was mainly to find a minimal attribute subset to keep the fuzzy positive region or the dependency degree unchanged. It needs to calculate the fuzzy equivalence or similarity classes, and calculate the lower approximation and dependency by using the inclusion relation between fuzzy classes and decision classes. The calculation of fuzzy classes is very time consuming, which is positively related to the number of attributes and the square of the number of objects. The kernel-based attribute reduction [5], [10], [42] was to map the low-dimensional inseparable data to the high-dimensional separable space, then select attributes. In the mapping process by using kernel functions, it consumes a lot of time to calculate the similarity between objects. The entropy-based attribute reduction [39]–[41] was to select a minimal attribute subset with the least uncertainty described by different entropies. It also needs to calculate equivalence or similarity classes. The neighborhood-based attribute reduction [3], [33], [34] consumes a lot of time to find a proper neighborhood parameter and calculate fuzzy classes.

An efficient and reasonable attribute evaluation function is helpful to improve classification performance and reduce search time. In this article, we establish the attribute evaluation function by directly considering the relationship between objects and decision categories, which avoid the cost of calculating the relationship between objects. In the first step, we define the fuzzy membership between objects and decision classes by minimizing the objective function (1), and compute the degree of aggregation (DA) of intraclass objects by the defined fuzzy membership. The degree of dispersion (DD) of between-class objects is defined by intraclass centers. In the second step, we define the separability of attribute subsets in fuzzy decision systems by combining DA and DD, in which the separability is used to characterize the classification ability of an attribute subset. In the last step, we design a heuristic algorithm to select attributes. Finally, we use the UCI [7], ELVIRA Biomedical [44], and handwritten digit datasets to verify the stability and effectiveness of the proposed algorithm by comparing it with some representative algorithms. From the distance-based evaluation index perspective, we choose the classic ReliefF-based attribute selection algorithm (RELIEF-F) [11], and the state-of-the art algorithms based on the neighborhood relation between objects, namely, the neighborhood discrimination index-based algorithm (HANDI) [1] and neighborhood mutual information-based feature selection algorithm (NMI) [3]. From the redundancy between features and the relevance between features and targets perspective, we choose the state-of-the art algorithm based on mutual information, namely, minimal-redundancy-maximal-relevance (mRMR) [45]. RELIEF-F updates the weights of attributes by using the difference between a sample and k nearest neighbors from the same class and the difference between the sample and k nearest neighbors from each of the different classes. The weights are updated repeatedly to evaluate the quality of attributes. HANDI describes the significance of attributes based on the distinguishing ability of the neighborhood similarity relation. NMI evaluates the significance of attribute subsets by joint neighborhood entropy between attributes and joint neighborhood entropy between attributes and decisions. mRMR selects feature subsets by minimizing the redundancy between features and maximizing the relevance between features and targets simultaneously. From the perspective of classification, the purpose of this article is to find an attribute subset that makes the intraclass objects compact and the between-class objects sparse. The experimental results show that the proposed algorithm has outstanding advantages in computational efficiency, classification accuracy, and the size of selected attributes.

This article is organized as follows. In Section II, we review some basic concepts of linear discriminant analysis (LDA) and analyze our research motivation. Then, the measures of DA and DD are defined to evaluate the significance of attribute subsets in Section III. In Section IV, a sequentially forward selection based on the separability (SFSS) algorithm is proposed to select attribute subsets. In Section V, we use public datasets to verify the feasibility, efficiency, and stability of SFSS. Finally, our conclusions and future work are elaborated in Section VI.

II. Related Work and Motivation

As we study attribute reduction in fuzzy decision systems, it is necessary to review the basic concepts of LDA [43] and fuzzy decision systems [13]. Then, we will introduce motivation for attribute reduction.

A. Linear Discriminant Analysis

Let $U = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$ be a training set with n samples. There are K categories, n_k is the number of samples of the kth category, $x_k^i \in R^m$ denotes the ith sample of the kth category. The aim of LDA is to find a projection vector, which can reduce the distance of samples from the same category and increase the distance of samples from different categories. The projection vector is obtained by using the following Fisher criterion:

$$v^* = \arg\max_v \frac{v^T S_b v}{v^T S_w v}$$

where S_b and S_w are the between-class and intraclass scatter matrices, respectively. They are calculated as follows:

$$S_b = \frac{1}{n}\sum_{k=1}^{K} n_k (c_k - c)(c_k - c)^T$$

$$S_w = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \left(x_k^i - c_k\right)\left(x_k^i - c_k\right)^T$$

where $c_k = [1/n_k]\sum_{i=1}^{n_k} x_k^i$ and $c = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_k^i$. The solution of the above projection vector can be transformed into

TABLE I
FUZZY DECISION SYSTEM

U	a_1	a_2	a_3	d	U	a_1	a_2	a_3	d
x_1	0.44	0.81	0.57	1	x_{10}	0.70	0.26	0.64	2
x_2	0.45	0.62	0.47	1	x_{11}	0.41	0.38	0.59	2
x_3	0.18	0.74	0.21	1	x_{12}	0.66	0.46	0.81	2
x_4	0.39	0.63	0.36	1	x_{13}	0.88	0.47	0.34	3
x_5	0.33	0.65	0.26	1	x_{14}	0.78	0.70	0.73	3
x_6	0.23	0.85	0.65	1	x_{15}	0.87	0.67	0.31	3
x_7	0.43	0.34	0.42	2	x_{16}	0.69	0.72	0.68	3
x_8	0.52	0.48	0.49	2	x_{17}	0.76	0.67	0.66	3
x_9	0.73	0.31	0.79	2	x_{18}	0.81	0.45	0.28	3

the following optimization problem:

$$v^* = \arg\max_{v^T v=1} v^T (S_w - \lambda S_b) v$$

where λ is a small positive constant. So we know that the optimal projection vector v is the eigenvector corresponding to the minimum eigenvalue of $S_w v = \lambda S_b v$.

LDA can map the data from high-dimensional space to low-dimensional space using linear transformation, which aims to improve the data separability. The attributes obtained by LDA transformation are not the children attributes of the original attributes, but are the mapping attributes in the new low-dimensional space. However, in many practical applications, we need to remove some redundant attributes from the original attributes without changing the meaning of attributes themselves.

B. Motivation

LDA uses the category information to find a projection that minimizes the distance of samples with the same category and maximizes the distance of samples from different categories so as to improve the classification accuracy. Inspired by this idea, we define two measures DA and DD, from the perspective of the intraclass compactness and the between-class sparsity, to select attributes without changing the original meaning of attributes.

We first introduce the basic concept of a fuzzy decision system. Let FDS $= (U, A, V, f)$ be a fuzzy decision system, where $U = \{x_1, x_2, \ldots, x_n\}$ is the universe of discourse, $A = C \cup D$ is the union of the conditional attribute set $C = \{a_1, a_2, \ldots, a_m\}$ and the decision attribute set $D = \{d_1, d_2, \ldots, d_s\}$, and $C \cap D = \varnothing$; $V = \cup_{a \in A} V_a$ is the attribute value domain. $f : U \times A \to V$ is a mapping, that is $\forall x \in U$ and $\forall a \in A$, we have $f(x, a) \in V_a$, where $f(x, a)$ represents the value of object x under the attribute a and $0 \le f(x, a) \le 1$. R is an equivalence relation which is formed by the conditional attribute C in the universe U. $[x]_R$ is an equivalence class of object x induced by the equivalence relation R, where $[x]_R = \{y \in U | f(x, a) = f(y, a)\ \forall a \in C\}$. U/C denotes the conditional partition of universe U under the conditional attribute C and U/D denotes the decision partition of universe U under the decision attribute D.

Example 1: A given fuzzy decision system is shown in Table I, where $U = \{x_1, x_2, \ldots, x_{18}\}$, $C = \{a_1, a_2, a_3\}$, and $D = \{d\}$. These objects are divided into three mutually exclusive classes by the decision attribute d. We show the distribution information of objects under different attribute subsets, as shown in Fig. 1. The red circles, green squares, and blue hexagonal stars represent classes with decision attribute values of 1, 2, and 3, respectively. Fig. 1(a) shows the distribution of objects under the attribute set $\{a_1, a_2, a_3\}$. Fig. 1(b)–(d) shows the distributions of objects under attribute subsets $\{a_1, a_2\}$, $\{a_1, a_3\}$, and $\{a_2, a_3\}$, respectively. Different attribute subsets have different distinguishing ability for objects. From Fig. 1, the separability of objects under the attribute subset $\{a_1, a_2\}$ is better than those of $\{a_1, a_3\}$ and $\{a_2, a_3\}$.

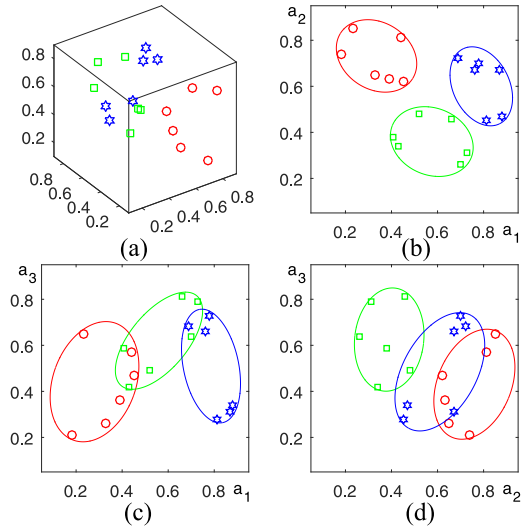

Fig. 1. Distributions of objects under different attribute subsets. (a) {a_1,a_2,a_3}. (b) {a_1,a_2}. (c) {a_1,a_3}. (d) {a_2,a_3}.

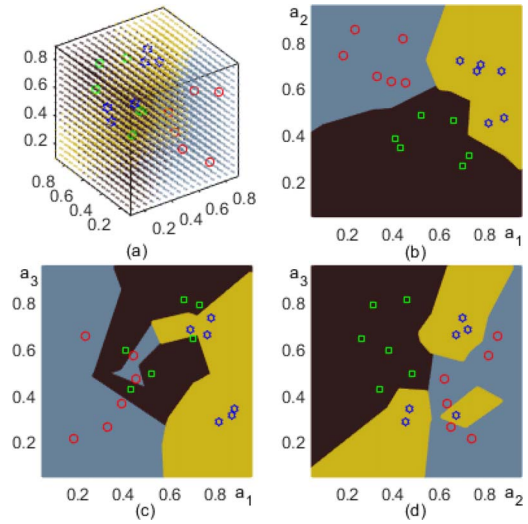

Fig. 2. Classification results of KNN under different attribute subsets. (a) {a_1,a_2,a_3}. (b) {a_1,a_2}. (c) {a_1,a_3}. (d) {a_2,a_3}.

Next, we use two classifiers to classify objects under different attribute sets. Figs. 2 and 3 are the classification results using KNN ($k = 2$) and RBF-SVM ($C = 1$ and $\sigma = 1$) classifiers, respectively. Obviously, the classification ability of KNN and RBF-SVM under the attribute subset $\{a_1, a_2\}$ is better than those of other attribute subsets. From Figs. 2 and 3, we can see that when the intraclass objects are closer and

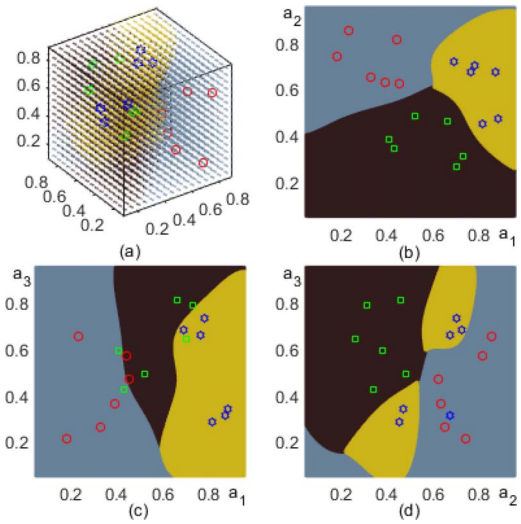

Fig. 3. Classification results of RBF-SVM under different attribute subsets. (a) {a1,a2,a3}. (b) {a1,a2}. (c) {a1,a3}. (d) {a2,a3}.

the between-class objects are more scattered, the classification results are better, and the overfitting can also be prevented. From the above example, we find that the closer the intraclass objects are and the more scattered the between-class objects are, the more conducive to classification. Thus, it is meaningful to select an attribute subset that is conducive to classification by studying the separability of attributes with respect to categories.

The purposes of attribute reduction for classification task, regression task and keeping the positive region of rough sets unchanged are different.

1) The purpose of attribute reduction of the classification task is to find a minimal attribute subset which can make the similarity of objects from same category higher and that of objects from different categories lower.
2) The purpose of attribute reduction of the regression task is to find a minimal attribute subset, so that when the similarity of objects under this attribute subset is high, the similarity of objects under decision attributes is also high.
3) The purpose of attribute reduction of keeping the positive region of rough set unchanged is to find a minimal attribute subset which can make the attribute information granule consistent or basically consistent with the decision information granule. In this article, we will reduce redundant and inconsistent attributes of the classification task based on the separability.

The main innovations of this article are as follows.

1) Considering that the distances, between objects and class centers, are negatively related to the fuzzy membership of the objects about the classes, we use fuzzy membership to describe the DA of intraclass objects.
2) The distances between class centers are used to measure the DD of between-class objects.
3) Based on the DA and the DD, we propose a separability measure to evaluate the significance of attribute subsets.
4) Directly considering the relation between objects and decision classes can greatly reduce the calculation time and improve the efficiency of reduction.

TABLE II
DESCRIPTION OF NOTATIONS

Notation	Description
FDS	A fuzzy decision system
U	A non-empty finite set of objects, called universe
D	A decision attribute set
U/D	The decision partition of U under D
D_k	The k^{th} decision class
C_k	The center of D_k under attribute subset B
μ_{ik}	The membership of x_i with respect to D_k under B
$d_B(x_i, D_k)$	The distance of x_i with respect to D_k under B
$DA_B(D_k)$	The degree of aggregation of D_k under B
$GDA_B(S)$	The degree of aggregation of FDS under B
$DD_B(S)$	The degree of dispersion of FDS under B
$DS_B(S)$	The separability of FDS under B
$SIG(a, B, D)$	The significance of a with respect to B under D

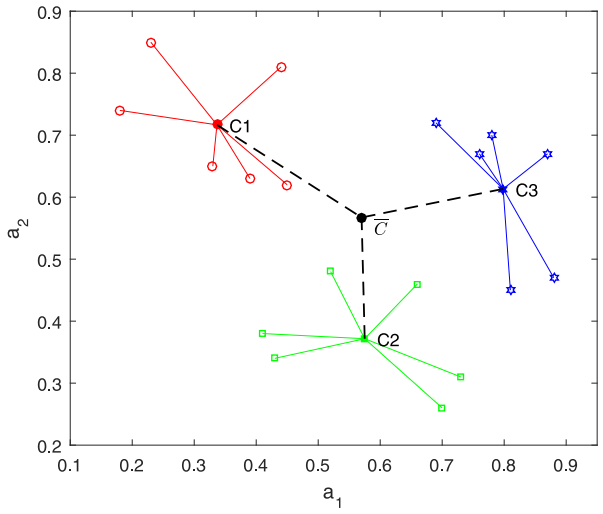

Fig. 4. Distribution of objects under attribute subset $\{a_1, a_2\}$.

5) Based on the separability measure and postpruning strategy, we design a sequentially forward selection algorithm for attribute reduction.

C. Notations

To facilitate reading and understanding, we first explain the notations to prepare for the following work. Detailed information is shown in Table II.

III. Measures of Intraclass Compactness and Between-Class Sparsity

In this section, we will first define the DA of intraclass objects by fuzzy membership and the DD of between-class objects by distance in the fuzzy decision system. Then, the rationality of the definition is discussed.

Let FDS $= (U, A, V, f)$ be a fuzzy decision system, where $U/D = \{D_1, D_2, \ldots, D_K\}$ and $D_k (k = 1, 2, \ldots, K)$ is the kth decision class. CP $= \{C_1, C_2, \ldots, C_K\}$ is a set of the center of each class under the attribute subset B, where $C_k = (c_k(a_1), c_k(a_2), \ldots, c_k(a_{|B|})), (k = 1, 2, \ldots, K, a_i \in B)$, and $c_k(a_i)$ is mean or median of all objects of the kth decision class under the attribute a_i. Fig. 4 shows the distribution of objects in Example 1 under attribute subset $\{a_1, a_2\}$, where C_1,

C_2, and C_3 are the mean centers of red circles, green squares, and blue hexagonal stars objects, respectively.

Let $\mu_B(x_i, D_k)$ be the membership of object x_i with respect to the decision class D_k under attribute subset B, abbreviated as μ_{ik}. Let $d_B(x_i, D_k)$ be the distance of object x_i with respect to D_k under B. In order to describe the relationship between the fuzzy membership $\mu_B(x_i, D_k)$ and the distance $d_B(x_i, D_k)$, we introduce the objective function

$$J = \sum_{k=1}^{K}\sum_{i=1}^{n} \mu_{ik}^m d_B^2(x_i, D_k) \tag{1}$$

where $0 \le \mu_{ik} \le 1 (i = 1, 2, \ldots, n \text{ and } k = 1, 2, \ldots, K)$, $\sum_{k=1}^{K} \mu_{ik} = 1 (i = 1, 2, \ldots, n)$, $m (m > 1)$ is a weighting exponent. When the objective function J obtains the minimum value, μ_{ik} has a negative correlation with $d_B(x_i, D_k)$. The smaller the distance is, the greater the membership is. d_B can take the Minkowski, Mahalanobis, cosine, and Hamming distances and so on. This article mainly introduces the Minkowski distance. The calculation method of $d_B(x_i, D_k)$ is

$$d_B(x_i, D_k) = \sqrt[p]{\sum_{a\in B} |f(x_i, a) - c_k(a)|^p} \tag{2}$$

where p can be any real number greater than or equal to 1. $d_B(x_i, D_k)$ is called the Manhattan distance if $p = 1$, the Euclidean distance if $p = 2$, and the Chebychev distance if $p = \infty$.

The objective function J is a constrained optimization problem. We use the Lagrange multiplier method to transform it into an unconstrained optimization problem. The unconstrained objective function is as follows:

$$F = \sum_{k=1}^{K}\sum_{i=1}^{n} \mu_{ik}^m d_B^2(x_i, D_k) + \sum_{i=1}^{n} \lambda_i \left(\sum_{k=1}^{K} \mu_{ik} - 1 \right) \tag{3}$$

where λ_i for $i = 1, 2, \ldots, n$ are Lagrange multipliers. The minimum value of (3) is required to find the stationary point of the function. Therefore, we calculate the partial derivatives of F with respect to μ_{ik} and λ_i, namely

$$\frac{\partial F}{\partial \mu_{ik}} = m\mu_{ik}^{m-1} d_B^2(x_i, D_k) + \lambda_i \tag{4}$$

$$\frac{\partial F}{\partial \lambda_i} = \sum_{k=1}^{K} \mu_{ik} - 1. \tag{5}$$

Let $[(\partial F)/(\partial \mu_{ik})] = 0$ and $[(\partial F)/(\partial \lambda_i)] = 0$, μ_{ik} is obtained as follows:

$$\begin{aligned} \mu_{ik} &= \left(\frac{-\lambda_i}{\mathrm{m} d_B^2(x_i, D_k)} \right)^{\frac{1}{m-1}} \\ &= \left(\frac{-\lambda_i}{m} \right)^{\frac{1}{m-1}} \left(\frac{1}{d_B(x_i, D_k)^{\frac{2}{m-1}}} \right). \end{aligned} \tag{6}$$

Based on (5) and (6), we can be further concluded

$$\begin{aligned} 1 = \sum_{k=1}^{K} \mu_{ik} &= \sum_{k=1}^{K} \left(\frac{-\lambda_i}{m} \right)^{\frac{1}{m-1}} \left(\frac{1}{d_B(x_i, D_k)^{\frac{2}{m-1}}} \right) \\ &= \left(\frac{-\lambda_i}{m} \right)^{\frac{1}{m-1}} \sum_{k=1}^{K} \frac{1}{d_B(x_i, D_k)^{\frac{2}{m-1}}}. \end{aligned} \tag{7}$$

$(-\lambda_i/m)^{(1/[m-1])}$ can be solved from (7), and bringing it into (6), we can obtain

$$\begin{aligned} \mu_{ik} &= \left(\frac{1}{\sum_{j=1}^{K} \frac{1}{d_B(x_i, D_j)^{\frac{2}{m-1}}}} \right) \left(\frac{1}{d_B(x_i, D_k)^{\frac{2}{m-1}}} \right) \\ &= \frac{1}{\sum_{j=1}^{K} \frac{d_B(x_i, D_k)^{\frac{2}{m-1}}}{d_B(x_i, D_j)^{\frac{2}{m-1}}}} = \frac{1}{\sum_{j=1}^{K} \left(\frac{d_B(x_i, D_k)}{d_B(x_i, D_j)} \right)^{\frac{2}{m-1}}}. \end{aligned} \tag{8}$$

From (8), we know that when m is equal to 2 or 3, the value of $(2/[m-1])$ is an integer and the solution of μ_{ik} is also related to d_B. If $p = 1$ in d_B and $m = 3$, we can reduce the square operation in μ_{ik} solving process. If $p = 2$ in d_B and $m = 2$, we can reduce the square root and square operation in μ_{ik} solving process. If p in d_B is equal to other values, we can obtain $m = (2/p) + 1$ to reduce the exponential operation in μ_{ik} solving process. We choose Euclidean distance in this article, so m is set to 2 to improve computational efficiency.

A new measure, called the DA of intraclass objects, is proposed to measure the closeness of intraclass objects by memberships of objects with respect to decision classes.

Definition 1: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $D_k \in U/D$ and the attribute subset $B \subseteq C$, the DA of the decision class D_k under B is defined as

$$\mathrm{DA}_B(D_k) = \frac{\sum_{x_i \in D_k} \mu_B(x_i, D_k)}{|D_k|} = \frac{\sum_{x_i \in D_k} \mu_{ik}}{|D_k|}. \tag{9}$$

Because the membership has a negative correlation with the distance, the larger DA_B is, the closer the intraclass objects are, under the attribute subset B.

Property 1: Let $B \subseteq C$ and $D_k \in U/D$, then $0 \le \mathrm{DA}_B(D_k) \le 1$.

Proof: $0 \le \mu_{ik} \le 1$, we can obtain $0 \le \sum_{x_i \in D_k} \mu_{ik} \le |D_k|$. Thus, $0 \le \mathrm{DA}_B(D_k) \le 1$. ■

Definition 2: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $U/D = \{D_1, D_2, \ldots, D_K\}$, $B \subseteq C$, the DA of the fuzzy decision system under B is defined as

$$\mathrm{GDA}_B(S) = \frac{\sum_{D_k \in U/D} \mathrm{DA}_B(D_k)}{|U/D|} \tag{10}$$

where S denotes the fuzzy decision system. $\mathrm{GDA}_B(S)$ is a measure of the intraclass compactness of all decision classes under B in the FDS.

Property 2: Let $B \subseteq C$ and $U/D = \{D_1, D_2, \ldots, D_K\}$, then $0 \le \mathrm{GDA}_B(S) \le 1$.

Proof: According to Property 1, $0 \le \mathrm{DA}_B(D_k) \le 1$, we can obtain $0 \le \sum_{D_k \in U/D} \mathrm{DA}_B(D_k) \le |U/D|$. Thus, $0 \le \mathrm{GDA}_B(S) \le 1$. ■

In addition to the compactness of the intraclass objects that can measure the separability of the system, the sparsity of the between-class objects is also an important index to measure the separability of the system. We will define the DD of between-class objects based on intraclass centers.

Let $\overline{C}$ be the center of all class centers under B, where $\overline{C} = (\overline{c}(a_1), \overline{c}(a_2), \ldots, \overline{c}(a_{|B|})), (a_i \in B)$, $\overline{c}(a_i) = (1/K) \sum_{i=1}^{K} c_k(a_i)$. $d_B(\overline{C}, C_k) = \sqrt{\sum_{a_i \in B} |\overline{c}(a_i) - c_k(a_i)|^2}$

denotes the distance from $\overline{C}$ to C_k under B. $\overline{C}$ is the center of C_1, C_2, and C_3 in Fig. 4. We will define the DD of a fuzzy decision system under B.

Definition 3: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $B \subseteq C$, the DD of the fuzzy decision system under B is defined as

$$\mathrm{DD}_B(S) = \frac{\sum_{k=1}^{K} d_B(\overline{C}, C_k)}{|U/D|} \tag{11}$$

where $\mathrm{DD}_B(S)$ is a sparsity measure of between-class objects under B. The larger $\mathrm{DD}_B(S)$ is, the more scattered the between-class objects are.

Property 3: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $B_1 \subseteq B_2 \subseteq C$, then $\mathrm{DD}_{B_1}(S) \leq \mathrm{DD}_{B_2}(S)$.

Proof: $B_1 \subseteq B_2$, hence, $d_{B_2}(\overline{C}, C_k) = \sqrt[2]{\sum_{a_i \in B_2} |\overline{c}(a_i) - c_k(a_i)|^2} = \sqrt[2]{\sum_{a_i \in B_1} |\overline{c}(a_i) - c_k(a_i)|^2 + \sum_{a_i \in B_2 - B_1} |\overline{c}(a_i) - c_k(a_i)|^2} \leq d_{B_1}(\overline{C}, C_k)$. Thus, $\mathrm{DD}_{B_1}(S) \leq \mathrm{DD}_{B_2}(S)$. ■

Property 4: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $B \subseteq C$, then $0 \leq \mathrm{DD}_B(S) \leq \sqrt{|B|}$.

Proof: FDS is a fuzzy decision system, so $0 \leq d_B(\overline{C}, C_k) \leq \sqrt{|B|}$, $0 \leq \sum_{k=1}^{K} d_B(\overline{C}, C_k) \leq K\sqrt{|B|}$, thus $0 \leq \mathrm{DD}_B(S) \leq \sqrt{|B|}$. ■

Remark 1: $\forall B \subseteq C$, we have $\mathrm{DD}_B(S) \leq \mathrm{DD}_C(S)$. According to Property 3, $\mathrm{DD}_B(S)$ is monotonically increasing with the size of B. Hence, when $B = C$, the DD of between-class objects is largest. $\forall B_1, B_2 \subseteq C$, when $|B_1| = |B_2|$, if $\mathrm{DD}_{B_1}(S) < \mathrm{DD}_{B_2}(S)$, the distinguishing ability of B_2 is greater than that of B_1 with between-class objects.

Definition 4: Let $\mathrm{GDA}_B(S)$ and $\mathrm{DD}_B(S)$ are the degrees of aggregation and dispersion of FDS under B, the separability of the fuzzy decision system under B is defined as

$$\mathrm{DS}_B(S) = \mathrm{GDA}_B(S) \cdot \mathrm{DD}_B(S). \tag{12}$$

When $B = \varnothing$, we set $\mathrm{DS}_B(S) = 0$. $\mathrm{DS}_B(S)$ is a measure, it describes the significance of the conditional attribute subset relative to the decision based on the intraclass compactness and between-class sparsity.

Property 5: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $\mathrm{DS}_B(S)$ is the separability of the FDS under B, then $0 \leq \mathrm{DS}_B(S) \leq \sqrt{|B|}$.

Proof: It follows directly from Properties 2 and 4. ■

To understand the above calculation process, we take $B = \{a_1, a_2\}$ to calculate DA, DD, and the separability of Example 1. The intraclass centers are $C_1 = (0.3367, 0.7167)$, $C_2 = (0.5750, 0.3717)$, and $C_3 = (0.7983, 0.6133)$. The distance $d_B(x_i, D_k)$ and the membership u_{ik} of object x_i with respect to D_k under B are shown in Table III. The gray cells in the Table III are the membership of intraclass objects. According to the values of the gray cells, we have $\mathrm{DA}_B(D_1) = 0.6327$, $\mathrm{DA}_B(D_2) = 0.5554$, and $\mathrm{DA}_B(D_3) = 0.6280$, thus

$$\mathrm{GDA}_B(S) = \frac{0.6327 + 0.5554 + 0.6280}{3} = 0.6054.$$

The center of C_1, C_2, and C_3 is $\overline{C} = (0.5700, 0.5672)$, thus, $d_B(\overline{C}, C_1) = 0.2771$, $d_B(\overline{C}, C_2) = 0.1956$, and $d_B(\overline{C}, C_3) = 0.2329$. According to (11), we compute

$$\mathrm{DD}_B(S) = \frac{0.2771 + 0.1956 + 0.2329}{3} = 0.2352.$$

Thus, we have

$$\mathrm{DS}_B(S) = 0.6054 \times 0.2352 = 0.1424.$$

TABLE III
DISTANCE AND FUZZY MEMBERSHIP UNDER $\{a_1, a_2\}$

U	$d_B(x_i, D_k)$			u_{ik}		
	D_1	D_2	D_3	D_1	D_2	D_3
x_1	0.1392	0.4587	0.4088	0.6082	0.1846	0.2072
x_2	0.1490	0.2780	0.3484	0.5093	0.2729	0.2178
x_3	0.1584	0.5401	0.6312	0.6476	0.1899	0.1625
x_4	0.1018	0.3177	0.4087	0.6372	0.2041	0.1587
x_5	0.0670	0.3708	0.4698	0.7557	0.1365	0.1078
x_6	0.1707	0.5898	0.6156	0.6382	0.1848	0.1770
x_7	0.3881	0.1484	0.4587	0.2242	0.5862	0.1897
x_8	0.2994	0.1215	0.3086	0.2255	0.5557	0.2188
x_9	0.5658	0.1668	0.3109	0.1610	0.5460	0.2930
x_{10}	0.5836	0.1676	0.3668	0.1647	0.5733	0.2620
x_{11}	0.3446	0.1652	0.4530	0.2600	0.5423	0.1977
x_{12}	0.4128	0.1226	0.2065	0.1571	0.5289	0.3140
x_{13}	0.5967	0.3205	0.1650	0.1543	0.2874	0.5583
x_{14}	0.4436	0.3871	0.0886	0.1398	0.1602	0.7000
x_{15}	0.5354	0.4196	0.0914	0.1229	0.1568	0.7202
x_{16}	0.3533	0.3668	0.1520	0.2332	0.2247	0.5421
x_{17}	0.4259	0.3510	0.0684	0.1185	0.1438	0.7377
x_{18}	0.5433	0.2477	0.1637	0.1536	0.3368	0.5096

TABLE IV
AGGREGATION, DISPERSION, AND SEPARABILITY UNDER DIFFERENT ATTRIBUTE SUBSETS

B	$GDA_B(S)$	$DD_B(S)$	$DS_B(S)$
$\{a_1\}$	0.6002	0.1556	0.0934
$\{a_2\}$	0.4563	0.1304	0.0595
$\{a_3\}$	0.3416	0.0726	0.0248
$\{a_1, a_2\}$	0.6054	0.2352	0.1424
$\{a_1, a_3\}$	0.4734	0.1965	0.0930
$\{a_2, a_3\}$	0.4534	0.1496	0.0679
$\{a_1, a_2, a_3\}$	0.5265	0.2500	0.1316

According to the above calculation process, we can calculate DA, DD, and the separability under different attribute subsets. The calculation results are shown in Table IV. From Table IV, we can see that the DA of intraclass objects under $\{a_1, a_2\}$ is larger than those of other attribute subsets. When $|B| = 1$, the DD of between-class objects under $\{a_1\}$ is larger than those of $\{a_2\}$ and $\{a_3\}$. When $|B| = 2$, the DD of between-class objects under $\{a_1, a_2\}$ is larger than those of $\{a_1, a_3\}$ and $\{a_2, a_3\}$. The DD of between-class objects under $\{a_1, a_2, a_3\}$ is the largest. The separability of FDS under $\{a_1, a_2\}$ is the largest. Thus, the best attribute subset is $\{a_1, a_2\}$, which is consistent with the visualization result of Fig. 1. In the following, we will give the definition of the optimal attribute subset.

Definition 5: Let FDS $= (U, A, V, f)$ be a fuzzy decision system, $B \subseteq C$, $a \in C - B$, $b \in B$. a is a redundant attribute for B, if $\mathrm{DS}_{B \cup \{a\}}(S) \leq \mathrm{DS}_B(S)$. b is an indispensable attribute in B, if $\mathrm{DS}_{B-\{b\}}(S) < \mathrm{DS}_B(S)$. B is a reduction of condition C with respect to decision D in FDS, iff B satisfies:

(1) $\mathrm{DS}_{B\cup\{a\}}(S) \leq \mathrm{DS}_B(S) \quad \forall a \in C - B$;

(2) $\mathrm{DS}_{B-\{b\}}(S) < \mathrm{DS}_B(S) \quad \forall b \in B$.

From Definition 5, we know that B is a reduction, if all the attributes in $C - B$ are redundant and all the attributes in B are indispensable, that is, B is an optimal attribute subset. It is a NP-hard problem to find the reduction of FDS, so we will design a heuristic algorithm to find a great informative and separable attribute subset.

IV. Attribute Reduction Algorithm Based on the Separability

Based on the above analysis, the proposed separability can be used to evaluate distinguishing ability of attribute subsets. The larger the separability is, the more distinguishable the attribute subset has. For a fuzzy decision system with N attributes, it is time consuming and even infeasible to calculate the separability of all candidate subsets ($2^N - 1$). At present, there are many common search strategies for attribute selection such as branch and bound (B&B), genetic algorithms (GA) and greedy selection (GS). The (B&B) algorithm has exponential complexity. The GA is an evolutionary algorithm, so it may have a certain degree of instability. The GS algorithm can quickly find an approximate optimal solution. Therefore, we choose the GS algorithm to search the optimal or suboptimal attribute subset. In view of the sequentially forward selection (SFS) and sequentially backward elimination (SBE) forms of the greedy algorithm, we design an attribute reduction algorithm based on the SFS.

First, we define the significance of an attribute relative to an attribute subset. Then we design a SFS attribute reduction algorithm as Algorithm 1.

Definition 6: Let $\mathrm{FDS} = (U, A, V, f)$ be a fuzzy decision system, $B \subseteq C$, $a \in C - B$, the significance of a relative to B is defined as

$$\mathrm{SIG}(a, B, D) = \mathrm{DS}_{B\cup\{a\}}(S) - \mathrm{DS}_B(S). \tag{13}$$

$\mathrm{SIG}(a, B, D)$ is a measure, which denotes the significance of attribute a with respect to B under decision D.

In Algorithm 1, *red* is initialized into an empty set and the time complexity is $O(1)$. In steps 3–5, this is a termination condition for feature selection, and the complexity is $O(1)$. In steps 6–9, we compute the significance of attribute a relative to *red*, and the complexity is $O(|C - red||U||U/D|)$. In step 11, attribute a_k with the maximum value of $\mathrm{SIG}(a_k, red, D)$ is selected, and the complexity is $O(1)$. In step 11, a_k is added into *red* and the complexity is $O(1)$. The worst search complexity is $O(|C|^2|U||U/D|)$ of Algorithm 1.

The termination parameter δ should be set in advance. How to set δ is a very important issue for attribute reduction. Stopping the search too early may lead to insufficient attributes used for learning; whereas stopping the search too late may lead to redundant attributes used for learning and overfitting may occur. People can set δ manually according to the needs of learning tasks. This article provides a search strategy of termination parameter δ based on the postpruning idea from [8]–[10]. Next, we will give the detailed search process. In the first step, the dataset is divided into two parts, one is the training set, the other is the validation set. In the second step, we select attributes by running the forward selection algorithm and record the order of attributes that have been selected for the training set, where the search range of δ is set to $[1, |C|]$ in low-dimensional data and that of δ is set to $[1, G]$ in high-dimensional data, $[1, G]$ is a given search range in advance. In the third step, we train classifiers on the training set under the selected sequential attribute subsets one by one, and record classification accuracies on the validation set. δ is set to the number of attributes with the largest average classification accuracies and the smallest number. Namely, in the kth time, we evaluate the first kth attributes which are selected in the second step with classification algorithms on the validation set.

Algorithm 1 SFSS

Input: A fuzzy decision system *FDS*.
Output: Attribute subset *red*.

1: Initialize: $red \leftarrow \varnothing$; // Initialization attribute subset is empty.
2: **while** $C - red \neq \varnothing$ **do**
3: **if** $|red| > \delta$ **then**
4: break; // Loop termination.
5: **end if**
6: **for** each $a_i \in C - red$ **do**
7: Compute the separability $DS_{B\cup\{a_i\}}(S)$;
8: Compute $SIG(a_i, red, D)$ according to Eq.(13);
9: **end for**
10: Find a_k with maximum value of $SIG(a_k, red, D)$;
11: $red \leftarrow a_k$; // Put coordinated a_k into *red*.
12: **end while**
13: return *red*;

V. Experimental Analysis

To verify the effectiveness and feasibility of the proposed algorithm (SFSS), we compare it with HANDI [1], NMI [3], RELIEF-F [11], and mRMR [45]. We compare them from three aspects: 1) the classification accuracies under different classifiers; 2) the number of the selected attributes; and 3) the running time of attribute reduction. All algorithms are executed in MATLAB 2015b and run in hardware environment with Inter Core i7-7700K @ 4.20 GHz, with 16-GB RAM.

We employ four well-known classifiers to estimate classification accuracies of these attribute reduction algorithms based on tenfold cross-validation. The four classifiers are k-nearest neighbor (KNN), radial basis function support vector machine (RBF-SVM), three layers fully connected neural network (FCNN), and random forest (RF). We set parameter $k = 3$ of KNN. The control term C and Gaussian kernel parameter σ of RBF-SVM are both one. Activation functions for the hidden layer and the output layer of FCNN are set to *sigmoid* and *softmax* functions, respectively. The number of neurons in the output layer is equal to the number of categories, and the number of the neurons of the hidden layer is equal to the mean of input and output layers. For RF, the number of decision trees is 20 and the number of variables

TABLE V
DATA DESCRIPTION

No.	Dataset	Sample	Attribute	Class
1	Wine	178	14	3
2	Wpbc	198	34	2
3	Seeds	210	8	3
4	Wdbc	569	31	2
5	Winequality-red	1599	12	6
6	Segmentation	2310	20	7
7	Spambase	4601	58	2
8	Winequality-white	4898	12	7
9	DLBCL-Stanford	47	4027	2
10	DLBCL-Harvard	77	7130	2
11	Lung-Cancer-1	181	12534	2
12	Lung-Cancer-2	203	12601	5

to select at random for each decision tree is the square root of the number of all variables. The characteristics of these four classifiers are that KNN does not have any parameters that need to be trained; SVM has a small number of parameters that need to be trained; FCNN has a large number of parameters that need to be trained; and RF has a small number of parameters that need to be trained and it integrates many weak classifiers to make its classification ability become stronger.

A. Experiment on UCI and ELVIRA Biomedical Datasets

The original data are normalized into the interval [0, 1] and are randomly divided into ten subsets. One is used as validation set and the remaining nine are used for training. After ten rounds, the median and fluctuation range of the classification accuracies are considered as the final performance. Twelve datasets from the UCI Machine-Learning Repository [7] and ELVIRA Biomedical Dataset Repository [44] are shown in Table V, where Lung-Cancer-1 is from the *Dana-Farber Cancer Institute* and Lung-Cancer-2 is from the *Brigham and Women's Hospital*.

Different datasets have different termination parameters for Algorithm 1. The termination parameters of all data are determined by the postpruning search strategy proposed in Section IV. The search ranges of δ on the first eight low-dimensional datasets are set to $[1, |C|]$, where $|C|$ is the number of conditional attributes. The search ranges of δ on the last four high-dimensional datasets are set to $[1, 50]$. Single accuracy and their average accuracy of four classifiers on *Wine*, *Wpbc* and *Lung-Cancer*-1 are shown in Figs. 5–7. Due to the space limitation, the search results for the remaining datasets are shown in the first nine figures of supplementary materials. The position of the red dot in each figure is the optimal average classification accuracy, and the corresponding number of attributes is the termination parameter. From these figures, we know that with the increase of the number of attributes, the average value of classification accuracies increases at the beginning, when the size reaches a certain degree, the average value of classification accuracies begins to decrease or stabilize.

Now, we analyze the trend of classification accuracy of each classifier after the red dot in Figs. 5–7. In the *Wine* dataset, the classification accuracies of KNN, FCNN, and

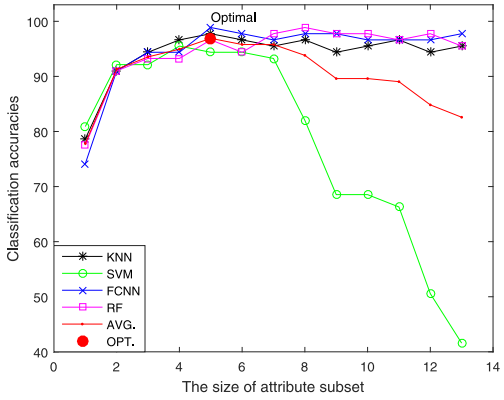

Fig. 5. Classification accuracies of Wine based on attribute ranking.

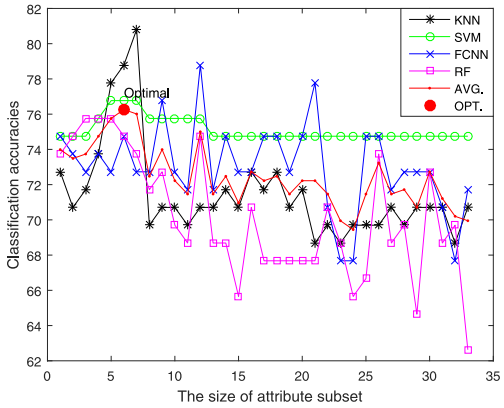

Fig. 6. Classification accuracies of Wpbc based on attribute ranking.

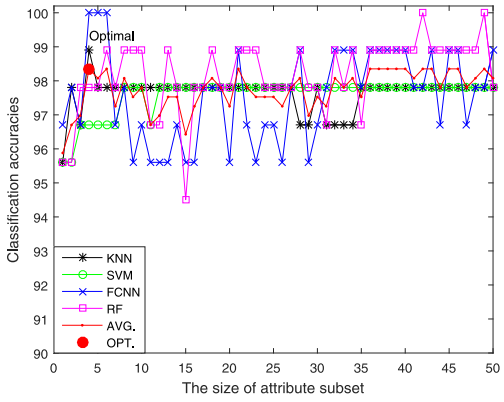

Fig. 7. Classification accuracies of Lung-Cancer-1 based on attribute ranking.

RF keep stable with the increase of attributes, while the classification accuracy of SVM is decreasing. This means that the attributes added after the red dot are redundant for KNN, FCNN, and RF, and inconsistent for SVM. In the *Wpbc* dataset, the classification accuracies of all classifiers are decreasing with the increase of attributes, which means that the redundant attributes are added after the red dot for four classifiers. In the *Lung-Cancer-1* dataset, the accuracies of KNN, FCNN, and RF are fluctuating with the increase of attributes, and the accuracy of SVM keeps stable. It shows that the added attributes are unstable for KNN, FCNN, and RF, and redundant for SVM. In this article, we choose the

TABLE VI
CLASSIFICATION ACCURACIES OF REDUCED DATA AND RANKS OF REDUCTION ALGORITHMS WITH KNN

Dataset	Raw data	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	91.667±8.333	91.667±8.333 (5)	94.118±5.882 (3.5)	94.444±5.556 (2)	94.118±5.882 (3.5)	97.222±2.778 (1)
Wpbc	68.684±21.316	76.579±13.421 (2)	72.500±17.500 (4)	68.421±21.053 (5)	75.000±10.000 (3)	81.579±18.421 (1)
Seeds	92.857±7.143	92.857±7.143 (2.5)	90.476±9.524 (5)	92.857±7.143 (2.5)	92.857±7.143 (2.5)	92.857±7.143 (2.5)
Wdbc	96.491±3.509	96.491±3.509 (1.5)	95.614±4.386 (4)	96.491±3.509 (1.5)	95.614±4.386 (4)	95.614±4.386 (4)
Winequality-red	60.938±5.937	59.688±8.437 (2)	59.062±5.313 (4)	58.125±8.750 (5)	59.375±6.875 (3)	61.250±5.000 (1)
Segmentation	95.671±1.299	96.537±1.299 (3.5)	96.537±0.866 (3.5)	97.186±1.082 (1)	95.022±2.814 (5)	96.753±1.515 (2)
Spambase	91.196±2.065	86.865±2.483 (5)	88.278±3.679 (1)	86.972±1.072 (3)	86.957±1.739 (4)	88.275±1.725 (2)
Winequality-white	58.265±2.959	56.633±4.184 (5)	59.694±3.571 (1)	58.367±2.653 (2.5)	58.367±2.653 (2.5)	57.755±3.673 (4)
DLBCL-Stanford	70.000±30.000	80.000±20.000 (4)	75.000±25.000 (5)	87.500±12.500 (2)	87.500±12.500 (2)	87.500±12.500 (2)
DLBCL-Harvard	85.714±14.286	81.250±18.750 (4)	81.250±18.750 (4)	87.500±12.500 (1)	81.250±18.750 (4)	85.714±14.286 (2)
Lung-Cancer-1	91.667±8.333	94.444±5.556 (3.5)	88.889±11.111 (5)	94.737±5.263 (2)	94.444±5.556 (3.5)	97.222±2.778 (1)
Lung-Cancer-2	90.000±5.000	90.476±9.524 (1)	87.500±7.500 (3.5)	87.500±7.500 (3.5)	77.500±17.500 (5)	90.000±5.000 (2)
Average	82.762±9.182	83.624±8.553 (3.250)	82.41±9.423 (3.625)	84.175±7.382 (2.583)	83.167±7.983 (3.500)	85.979±6.600 (2.042)

TABLE VII
CLASSIFICATION ACCURACIES OF REDUCED DATA AND RANKS OF REDUCTION ALGORITHMS WITH SVM

Dataset	Raw data	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	45.098±21.569	91.667±8.333 (2)	86.111±8.333 (4)	91.176±8.824 (3)	83.333±11.111 (5)	94.444±5.556 (1)
Wpbc	77.368±17.368	84.211±15.789 (1)	77.368±17.368 (4.5)	77.368±17.368 (4.5)	82.368±12.368 (3)	82.500±17.500 (2)
Seeds	90.476±9.524	90.476±9.524 (2.5)	88.095±7.143 (5)	90.476±9.524 (2.5)	90.476±9.524 (2.5)	90.476±9.524 (2.5)
Wdbc	64.912±10.526	94.737±5.263 (2.5)	91.228±8.772 (5)	94.737±3.509 (2.5)	92.982±5.263 (4)	95.614±4.386 (1)
Winequality-red	65.000±8.750	65.625±6.875 (3)	64.063±7.813 (4)	67.188±6.563 (1.5)	61.250±7.500 (5)	67.188±8.437 (1.5)
Segmentation	84.416±4.762	95.238±3.030 (2)	88.312±4.762 (5)	95.022±3.680 (3)	93.290±2.381 (4)	95.455±1.515 (1)
Spambase	77.935±2.065	87.609±3.261 (4)	91.413±3.370 (1)	89.457±1.848 (3)	87.187±1.100 (5)	89.674±2.500 (2)
Winequality-white	61.327±4.388	58.219±4.026 (5)	61.224±3.265 (4)	61.794±3.716 (2)	62.755±2.551 (1)	61.290±2.923 (3)
DLBCL-Stanford	35.000±15.000	50.000±30.000 (2.5)	40.000±40.000 (4)	50.000±50.000 (2.5)	35.000±15.000 (5)	60.000±40.000 (1)
DLBCL-Harvard	75.000±25.000	62.500±37.500 (3.5)	62.500±37.500 (3.5)	62.500±37.500 (3.5)	62.500±37.500 (3.5)	72.321±15.179 (1)
Lung-Cancer-1	86.111±8.333	88.889±11.111 (3)	83.333±16.667 (4.5)	97.222±2.778 (1.5)	83.333±11.111 (4.5)	97.222±2.778 (1.5)
Lung-Cancer-2	63.810±16.190	63.810±16.190 (5)	67.500±12.500 (3)	67.500±12.500 (3)	67.500±12.500 (3)	68.690±16.310 (1)
Average	68.871±11.956	77.748±12.575 (3.000)	75.096±13.958 (3.958)	78.703±13.151 (2.708)	75.165±10.659 (3.792)	81.240±10.551 (1.542)

TABLE VIII
CLASSIFICATION ACCURACIES OF REDUCED DATA AND RANKS OF REDUCTION ALGORITHMS WITH FCNN

Dataset	Raw data	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	94.444±5.556	88.889±11.111 (5)	91.667±8.333 (3.5)	91.667±8.333 (3.5)	97.059±2.941 (2)	97.222±2.778 (1)
Wpbc	76.579±13.421	80.000±20.000 (4)	81.711±13.289 (2)	68.816±16.184 (5)	81.579±18.421 (3)	82.500±17.500 (1)
Seeds	92.857±7.143	92.857±7.143 (2)	90.476±4.762 (4)	90.476±4.762 (4)	90.476±4.762 (4)	95.238±4.762 (1)
Wdbc	96.491±3.509	96.491±3.509 (2)	94.737±5.263 (5)	96.491±3.509 (2)	95.614±4.386 (4)	96.491±1.754 (2)
Winequality-red	57.500±6.250	59.375±5.000 (2)	58.438±5.312 (4)	58.750±6.250 (3)	55.937±8.438 (5)	60.313±4.688 (1)
Segmentation	94.589±1.515	94.589±1.948 (3)	95.238±1.299 (1)	88.961±7.576 (5)	94.156±1.948 (4)	94.805±2.165 (2)
Spambase	88.804±2.500	82.391±6.304 (3)	86.422±5.118 (2)	80.652±8.913 (5)	80.996±2.518 (4)	89.239±3.370 (1)
Winequality-white	52.551±4.184	53.936±3.732 (5)	54.856±3.631 (3)	55.060±3.836 (2)	53.980±2.959 (4)	56.339±4.192 (1)
DLBCL-Stanford	50.000±30.000	80.000±20.000 (4)	87.500±12.500 (2)	70.000±30.000 (5)	87.500±12.500 (2)	87.500±12.500 (2)
DLBCL-Harvard	75.000±25.000	87.500±12.500 (2)	85.714±14.286 (4)	87.500±12.500 (2)	81.250±18.750 (5)	87.500±12.500 (2)
Lung-Cancer-1	86.111±8.333	94.444±5.556 (3)	88.889±11.111 (5)	94.444±5.556 (3)	94.444±5.556 (3)	97.222±2.778 (1)
Lung-Cancer-2	63.810±16.190	80.952±14.286 (3)	85.000±10.000 (2)	80.238±10.238 (4)	77.500±17.500 (5)	85.714±14.286 (1)
Average	77.395±10.300	82.619±9.257 (3.167)	83.387±7.909 (3.125)	80.255±9.805 (3.625)	82.541±8.390 (3.750)	85.840±6.939 (1.333)

number of attributes with the best average classification accuracy as the termination parameter. For a specific classification task, the number of attributes with the best classification accuracy under the selected classifier can be set to the termination parameter. For the classifier with good generalization ability, we can add some attributes appropriately to improve the robustness of the classifier. For the classifier with poor generalization ability, it is recommended to select the number of attributes with the best classification performance on the validation set.

The related parameters of HANDI and NMI are searched by Wang *et al.* [1] and Hu *et al.* [3], respectively. The termination parameters of RELIEF-F and mRMR are searched by the same method as SFSS. The termination parameters of RELIEF-F, mRMR and SFSS are the same as the number of attributes retained after attribute reduction. The detailed results are shown in Table X.

The classification accuracy is an important index to evaluate the performances of attribute reduction algorithms. Under the four classifiers, the median and fluctuation range of classification accuracies of the reduced datasets based on five algorithms and the raw data are presented in Tables VI–IX, where the values in brackets are the ranks of reduction algorithms. The underlined symbols represent the highest classification accuracies.

From Tables VI–IX, under the four classifiers, the average classification accuracies of HANDI, RELIEF-F, mRMR, and SFSS are all better than those of the raw data. The

TABLE IX
CLASSIFICATION ACCURACIES OF REDUCED DATA AND RANKS OF REDUCTION ALGORITHMS WITH RF

Dataset	Raw data	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	97.059±2.941	91.667±8.333 (5)	94.444±5.556 (3)	94.118±5.882 (4)	97.059±2.941 (2)	98.889±11.111 (1)
Wpbc	75.000±15.000	77.368±17.368 (5)	80.000±15.000 (2)	79.868±14.868 (3.5)	79.868±14.868 (3.5)	82.500±12.500 (1)
Seeds	90.476±9.524	92.857±7.143 (3.5)	92.857±7.143 (3.5)	95.238±4.762 (1)	92.857±7.143 (3.5)	92.857±7.143 (3.5)
Wdbc	95.614±4.386	96.491±3.509 (2)	96.491±3.509 (2)	95.614±2.632 (4)	96.491±3.509 (2)	93.860±4.386 (5)
Winequality-red	67.734±7.109	66.563±7.188 (3)	69.375±6.250 (2)	65.938±5.938 (5)	66.250±6.875 (4)	71.875±8.125 (1)
Segmentation	97.835±1.299	97.835±0.866 (3.5)	98.052±1.082 (2)	97.835±1.299 (3.5)	97.619±1.515 (5)	98.320±1.515 (1)
Spambase	94.783±1.087	91.967±1.946 (4)	93.913±1.739 (1)	92.283±1.196 (3)	89.239±2.283 (5)	93.043±1.304 (2)
Winequality-white	68.849±2.930	67.925±1.463 (5)	68.844±1.973 (2)	68.333±1.667 (3)	68.265±2.143 (4)	69.828±3.542 (1)
DLBCL-Stanford	80.000±20.000	87.500±12.500 (3)	75.000±25.000 (5)	80.000±20.000 (4)	90.000±10.000 (1.5)	90.000±10.000 (1.5)
DLBCL-Harvard	75.000±25.000	81.250±18.750 (3.5)	87.500±12.500 (1.5)	81.250±18.750 (3.5)	75.000±25.000 (5)	87.500±12.500 (1.5)
Lung-Cancer-1	97.222±2.778	94.444±5.556 (3.5)	94.444±5.556 (3.5)	94.737±5.263 (2)	88.889±11.111 (5)	97.222±2.778 (1)
Lung-Cancer-2	88.095±11.905	88.095±11.905 (2)	87.500±12.500 (3.5)	85.000±5.000 (5)	87.500±12.500 (3.5)	92.500±7.500 (1)
Average	85.639±8.663	86.164±8.044 (3.583)	86.535±8.151 (2.583)	85.851±7.271 (3.458)	85.753±8.324 (3.667)	89.033±6.867 (1.708)

TABLE X
AVERAGE SIZE OF SELECTED ATTRIBUTE SUBSETS

Dataset	Raw data	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	13	6.1	7.1	6	7	5
Wpbc	33	5.7	7.7	17	7	6
Seeds	7	5.2	4	7	5	4
Wdbc	30	10.9	9.7	8	8	9
Winequality-red	11	7.5	7.9	10	5	7
Segmentation	19	7.9	13.2	8	10	8
Spambase	57	14.5	13.6	11	11	12
Winequality-white	11	5.8	7.6	7	8	8
DLBCL-Stanford	4026	11.1	10.3	10	22	15
DLBCL-Harvard	7129	14.3	10.8	20	14	15
Lung-Cancer-1	12533	6.7	6	9	20	4
Lung-Cancer-2	12600	16.9	15.3	38	11	9
Average	3039.083	9.383	9.433	12.583	10.667	8.5

performances of NMI on SVM, FCNN, and RF classifiers are better than those of the raw data, and its performance on KNN classifier is slightly worse than that of the raw data. Moreover, SFSS performs better than the other four methods and raw data from the average classification accuracy and the average fluctuation perspectives. The average classification accuracies of SFSS under KNN, SVM, FCNN, and RF are 3.257%, 12.369%, 8.445%, and 3.394% higher than those of raw data, respectively. Out of the 48 cases, HANDI, NMI, RELIEF-F, mRMR, and SFSS achieve the highest classification accuracies in 8, 8, 11, 7, and 35, respectively. Obviously, the performances of the first four reduction algorithms are comparable. The performance of SFSS is obviously better than the other four algorithms. From the results under the four classifiers, we can see that SFSS is very effective and robust for attribute reduction of classification task.

Next, we use the Friedman test [48] to evaluate the significant difference of the aforementioned five algorithms in the performance of attribute reduction. We first examine whether there are significant differences between the five algorithms on these datasets. The Friedman statistic is defined as $\chi_F^2 = (12N/[k(k+1)])(\sum_{i=1}^{k} R_i^2 - ([k(k+1)^2]/4))$ and $F = ([(N-1)\chi_F^2]/[N(k-1)-\chi_F^2])$, where N is the number of datasets, k is the number of algorithms, and R_i is the average rank of algorithm i in all the datasets. F follows a Fisher distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. The critical value of F distribution at the significance level α can be obtained by calling the subprogram $icdf('F', 1-\alpha, k-1, (k-1)(N-1))$ in MATLAB 2015b. Hence, we obtain the critical value $F(4, 44) = 2.0772$ when $\alpha = 0.1$. According to the ranks of the algorithms in Tables VI–IX, we can obtain that $F = 2.4061$ for KNN, $F = 6.6171$ for SVM, $F = 6.6722$ for FCNN, and $F = 4.3453$ for RF. All four values of F are greater than the critical value $F(4, 44)$, so we reject the null-hypothesis. Therefore, the five algorithms are significantly different under the four classifiers.

Furthermore, we use a *post-hoc* test such as the Bonferroni–Dunn test [47] to explore which algorithms are different in statistical terms. The performance of two algorithms is regarded as being significantly different if the distance of the average ranks of the two algorithms exceeds the critical distance $CD_\alpha = q_\alpha\sqrt{([k(k+1)]/6N)}$, where q_α is the critical tabulated value for the test. From [47], when $k = 5$ and $\alpha = 0.1$, $q_{0.1} = 2.241$. So $CD_{0.1} = 1.4466$. From Table VI, the distances of average ranks of SFSS to NMI and mRMR are greater than 1.4466 for KNN. Thus, the performance of SFSS is significantly better than those of NMI and mRMR at $\alpha = 0.1$. However, the Bonferroni–Dunn test is not powerful enough to detect any significant differences among SFSS, HANDI, and RELIEF-F. According to Table VII, SFSS performs significantly better than HANDI, NMI, and mRMR for SVM. However, we do not have sufficient evidence to show that there is a significant difference between SFSS and RELIEF-F. From Table VIII, we can obtain that SFSS performs significantly better than the other four algorithms for FCNN. From Table IX, SFSS performs significantly better than HANDI, RELIEF-F, and mRMR for RF. We do not have sufficient evidence to show that there is a significant difference between SFSS and NMI by Bonferroni–Dunn test. In summary, SFSS performs better than the other algorithms in most cases. At the same time, we use the Nemenyi test graph [47] to show the statistical difference between SFSS and the other four algorithms. We connect the groups of algorithms that are not significantly different (Fig. 8) at $p = 0.1$. We also show the critical difference ($CD = 1.5873$) above the graph, where $k = 5$ and $\alpha = 0.1$, $q_{0.1} = 2.459$. From Fig. 8, we can conclude that the Nemenyi test is not powerful enough to detect any significant differences among HANDI, NMI, RELIEF-F, and mRMR. The performance of SFSS is significantly better than that of HANDI and RELIEF-F under classifiers FCNN

TABLE XI
RUNNING TIME OF REDUCTION WITH DIFFERENT ALGORITHMS (S)

Dataset	HANDI	NMI	RELIEF-F	mRMR	SFSS
Wine	0.0862±0.0138	0.1267±0.0083	0.0899±0.0075	0.0947±0.0534	0.0216±0.0189
Wpbc	0.2424±0.1398	0.4815±0.0598	0.1573±0.0361	0.0448±0.0019	0.0070±0.0002
Seeds	0.0420±0.0089	0.0528±0.0092	0.0619±0.0106	0.0427±0.0024	0.0012±0.0000
Wdbc	2.2324±0.0897	2.4547±0.1712	0.4684±0.1345	0.0462±0.0023	0.0139±0.0006
Winequality-red	2.9663±0.1264	3.0638±0.1134	0.8057±0.047	0.0443±0.0030	0.0158±0.0008
Segmentation	13.3581±0.9059	20.4595±1.6067	1.5509±0.1835	0.0477±0.0028	0.0544±0.0017
Spambase	371.5869±19.3123	365.2280±18.8375	13.7437±1.8206	0.0840±0.0023	0.2274±0.0020
Winequality-white	25.0813±1.9759	27.4469±1.0773	3.6597±0.8241	0.0470±0.0029	0.0502±0.0012
DLBCL-Stanford	10.7185±1.1090	15.1395±1.4596	1.4216±0.0698	1.1469±0.0699	1.0238±0.0665
DLBCL-Harvard	48.6827±8.2103	51.2846±5.4096	4.4353±0.3422	1.2911±0.0266	1.5354±0.0184
Lung-Cancer-1	105.7291±11.6088	136.9752±22.7660	22.1078±2.7800	3.4996±0.1290	1.6776±0.0239
Lung-Cancer-2	381.7024±32.8598	435.4320±35.5697	24.1460±2.3923	2.9360±0.0253	2.4086±0.0230
Average	80.2023±6.3634	88.1788±7.2574	6.054±0.7207	0.7771±0.0268	0.5864±0.0131

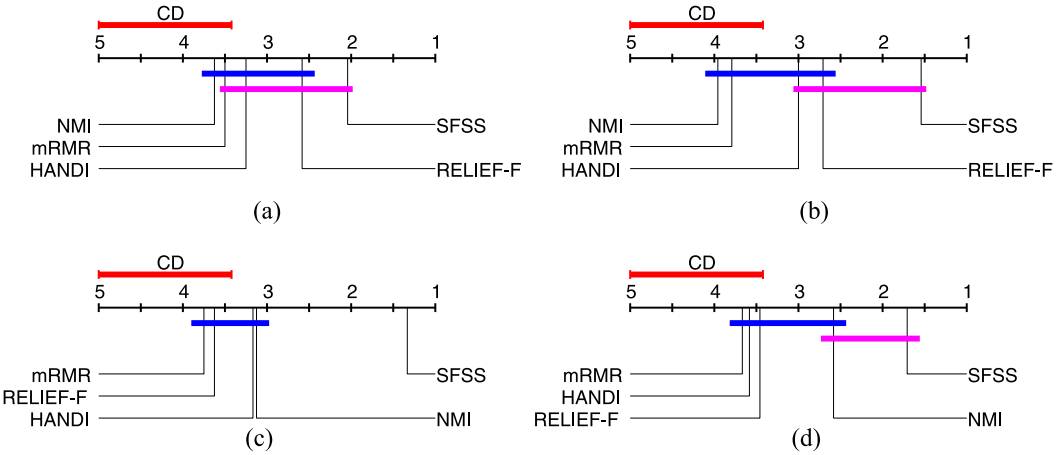

Fig. 8. Comparison of all reduction algorithms against each other with the Nemenyi test. Groups of algorithms that are not significantly different (at $p = 0.1$) are connected. Comparison of all reduction algorithms under (a) KNN, (b) SVM, (c) FCNN, and (d) RF.

and RF. The performance of SFSS is significantly better than that of NMI under classifiers KNN, SVM, and FCNN. The performance of SFSS is significantly better than that of mRMR under classifiers SVM, FCNN, and RF.

The size of the attribute subset, which is selected by the attribute reduction algorithms, is another important evaluation index for attribute reduction. We aim to find attribute subsets with high classification accuracy and as few attributes as possible. The average size of the selected attribute subsets with tenfold cross-validation is shown in Table X. From Table X, the average size of the selected attribute subsets by SFSS (8.5) is minimal compared with those of the other methods. In the case of ensuring classification accuracy, SFSS greatly reduces the dimension of the raw data.

In attribute reduction, the running time is still an important index to measure the feasibility and effectiveness of an algorithm. The median and fluctuation of running time based on tenfold cross-validation of five algorithms are shown in Table XI. From Table XI, we can find that the performance of SFSS is significantly better than those of HANDI and NMI. The performance of SFSS is slightly better than that of RELIEF-F. The performance of SFSS is comparable to that of mRMR. From the computing speed view, the performance of SFSS is efficient and robust.

In order to visualize the raw data and the reduced data selected by reduction algorithms, we use t-SNE [12] to visualize the raw data and reduced data in a 2-D map. Due to the space limitation, the results of the first eight datasets are shown in Figs. 9 and 10, and the results of the last four datasets are shown in the last figure of supplementary materials. From these figures, we can see that the separability of SFSS outperforms those of the other algorithms in the most cases. In the 2-D map, when the class overlap of the raw data (*Wine*, *Seeds*, *Wdbc*, *Segmentation*, *Spambase*, *DLBCL-Stanford*, *DLBCL-Harvard*, and *Lung-Cancer*-1) is less, SFSS can well separate different categories. In this case, the intraclass aggregation and the between-class dispersion for reduced data by using SFSS are larger than those of others. When the class overlap of the raw data (*Wpbc*, *Winequality-red*, *Winequality-white*, and *Lung-Cancer*-2) is relatively large, the intraclass aggregation and the between-class dispersion for reduced data by using SFSS have some improvements over the others.

B. Experiment on MNIST Database of Handwritten Digits

The MNIST[1] database is a well-known database for handwritten digit recognition. It has a training set of 60 000 images

[1]http://yann.lecun.com/exdb/mnist/

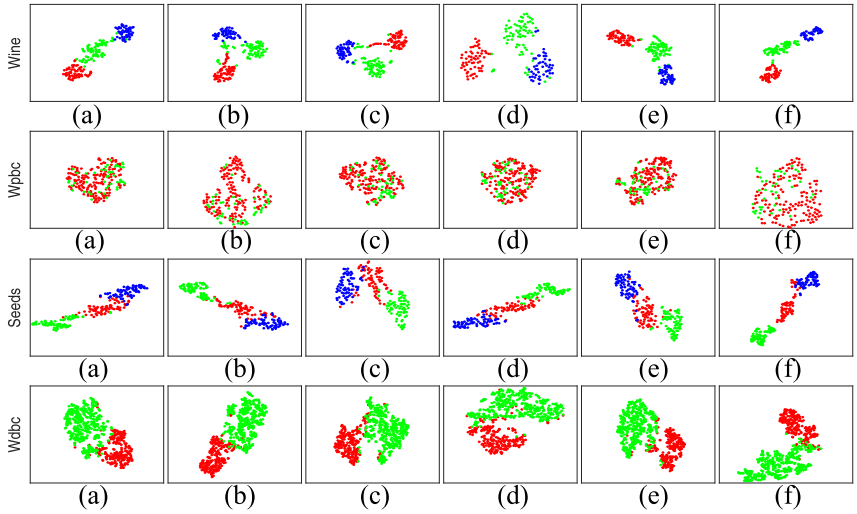

Fig. 9. t-SNE visualization of different algorithms in the first four datasets. (a) RAW. (b) HANDI. (c) NMI. (d) RELIEF-F. (e) mRMR. (f) SFSS.

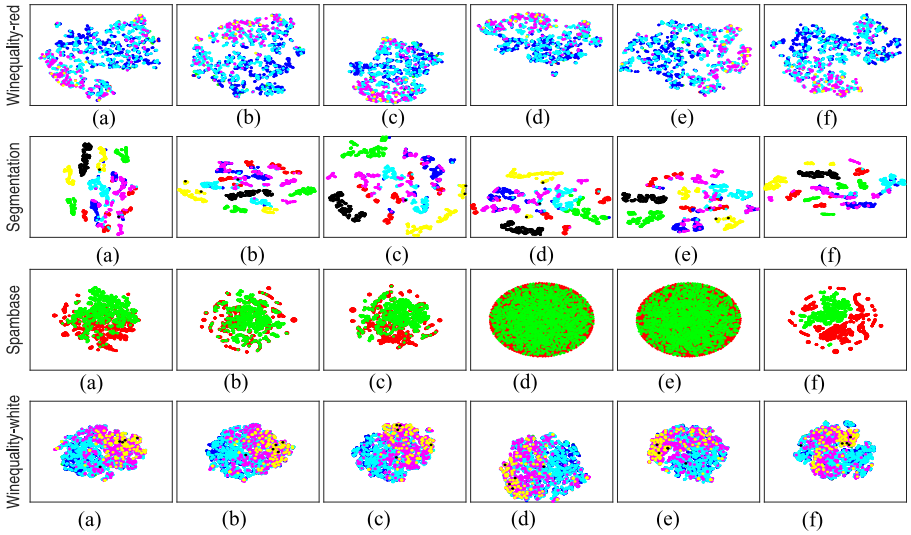

Fig. 10. t-SNE visualization of different algorithms in the middle four datasets. (a) RAW. (b) HANDI. (c) NMI. (d) RELIEF-F. (e) mRMR. (f) SFSS.

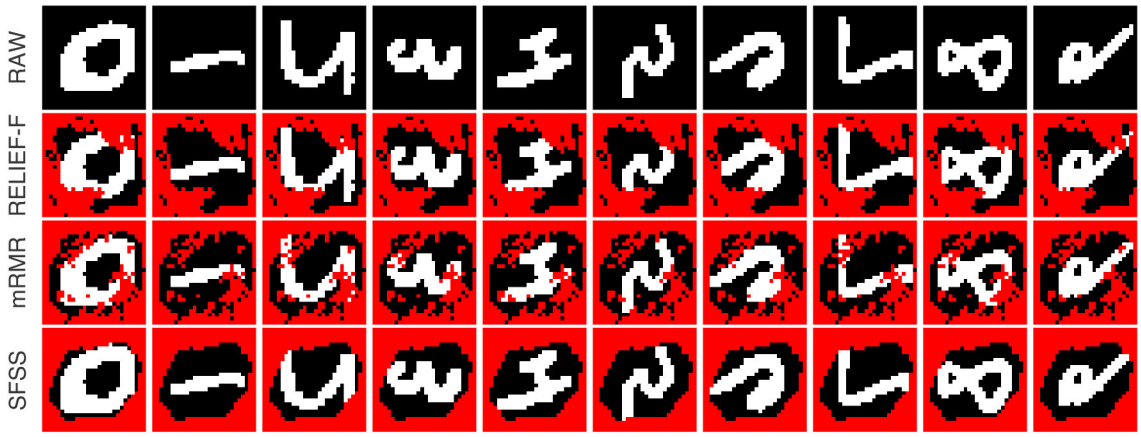

Fig. 11. Results of reduction algorithms in the MNIST dataset.

and a test set of 10 000 images, and each image has 28×28 pixels. The MNIST has ten categories, namely, digits 0–9. In order to observe the distributions of the selected attributes on images, we use reduction algorithms to retain half pixels and remove half pixels for each image. For a given neighborhood radius, when the number of attributes increases to a certain extent, the size of the neighborhood granule will not change any more. Therefore, HANDI and NMI are not suitable

TABLE XII
ACCURACIES OF VALIDATION SET AND TEST SET IN THE MNIST DATASET

Classifiers		Average			Best		
		RELIEF-F	mRMR	SFSS	RELIEF-F	mRMR	SFSS
KNN	val.	97.312	97.173	97.307	97.375	97.300	97.433
	test	96.862	97.010	96.936	96.960	97.120	97.020
SVM	val.	94.657	93.937	94.792	94.958	94.050	95.025
	test	94.972	94.352	95.196	95.120	94.390	95.240
FCNN	val.	96.455	96.210	96.555	96.675	96.317	96.650
	tes	96.702	96.382	96.708	96.790	96.640	96.820
RF	val.	95.705	95.380	95.718	96.025	95.783	95.875
	test	95.722	95.426	95.808	95.910	95.660	95.970

to retain large-scale attributes for attribute reduction. So we reduce the attributes of MNIST by RELIEF-F, mRMR, and SFSS. We use 10×5 cross-validation to select attributes on the training set of 60 000 images. The test set of 10 000 images is used as independent test data. In order to observe the removed pixels and retained pixels simultaneously, we select one image from each category to show the reduction results. The details are shown in Fig. 11, where red pixels are the removed pixels by attribute selection algorithms. MNIST handwritten digit images are manually collected, where the useful information is mainly concentrated in the middle of each image, and the edges are mostly filled information. In Fig. 11, from the distribution of red pixels in the edge of each image, we can see that all three algorithms delete many filled pixels. Moreover, SFSS performs better than RELIEF-F and mRMR. By observing the distribution of red pixels in the middle of each image, we can find that RELIEF-F and mRMR both delete some valuable pixels that are useful for digit recognition (such as digits 0, 2, 4, 5, 8). However, for SFSS, the pixels with distinguishing ability for digit recognition are preserved more completely. In terms of the continuity of the preserved regions, RELIEF-F and mRMR have poor continuity, while SFSS has good continuity. Therefore, the relevance of the selected pixels by SFSS is high for digit recognition. Table XII presents the average and best classification performances of the selected attributes by RELIEF-F, mRMR, and SFSS on validation data and independent test data of MNIST. The underlined symbols represent the highest classification accuracies. As can be seen from Table XII, the three algorithms are effective for MNIST, and the mean of all classification accuracies is 96.0496% and the variance is 0.0398%. The performances of the three algorithms are comparable. In most cases, the performance of SFSS is slightly better than those of the other two algorithms.

VI. CONCLUSION

Removing redundant attributes is an important step before performing classification and regression learning. It can decrease the cost of learning and improve learning performance. In this article, we have proposed an attribute reduction algorithm based on the aggregation of intraclass objects and the dispersion of between-class objects for fuzzy decision systems. The aggregation of intraclass objects and the dispersion of between-class objects were considered simultaneously to measure the significance of attribute subsets in fuzzy decision systems. Then, a postpruning strategy was introduced to search the termination parameter and prevent overfitting. Twelve public datasets from UCI and ELVIRA Biomedical repositories and MNIST handwritten digits were used to compare the performance of SFSS with those of classical algorithms. Experimental analysis and statistical test showed that SFSS can fast find a small and effective attribute subset and obtain high classification performance.

This article mainly studies the decision systems with real-valued attributes. In the future, we will study the attribute selection model of heterogeneous decision systems, which have category, real-valued, and interval-valued attributes simultaneously.

REFERENCES

[1] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong, "Feature selection based on neighborhood discrimination index," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2986–2999, Jul. 2018.

[2] C. Wang, Y. Huang, M. Shao, Q. Hu, and D. Chen, "Feature selection based on neighborhood self-information," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4031–4042, Sep. 2020, doi: 10.1109/TCYB.2019.2923430.

[3] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, and W. Pedrycz, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," *Expert Syst. Appl.*, vol. 38, pp. 10737–10750, Sep. 2011.

[4] Q. Hu, W. Pedrycz, D. Yu, and J. Lang, "Selecting discrete and continuous features based on neighborhood decision error minimization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 137–150, Feb. 2010.

[5] L. Chen, D. Chen, and H. Wang, "Fuzzy kernel alignment with application to attribute reduction of heterogeneous data," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 7, pp. 1469–1478, Jul. 2019.

[6] A. Mariello and R. Battiti, "Feature selection based on the neighborhood entropy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6313–6322, Dec. 2018.

[7] D. Dua and C. Graff, *UCI Machine Learning Repository*, Dept. School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[8] M. Bohanec and I. Bratko, "Trading accuracy for simplicity in decision trees," *Mach. Learn.*, vol. 15, pp. 223–250, Jun. 1994.

[9] F. Esposito, D. Malerba, and G. Semeraro, "A comparative analysis of methods for pruning decision trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 476–491, May 1997.

[10] Q. Hu, D. Yu, W. Pedrycz, and D. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, Nov. 2011.

[11] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of reliefF and RReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, Oct. 2003.

[12] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[13] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[14] Y. Guo, E. C. C. Tsang, W. Xu, and D. Chen, "Local logical disjunction double-quantitative rough sets," *Inf. Sci.*, vol. 500, pp. 87–112, Oct. 2019.

[15] Y. Guo *et al.*, "Incremental updating approximations for double-quantitative decision-theoretic rough sets with the variation of objects," *Knowl. Based Syst.*, vol. 189, Feb. 2020, Art. no. 105082. [Online]. Available: https://doi.org/10.1016/j.knosys.2019.105082

[16] E. C. C. Tsang, D. G. Chen, D. S. Yeung, X.-Z. Wang, and J. W. T. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, Oct. 2008.

[17] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, "Selection of rich model steganalysis features based on decision rough set α-positive region reduction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 336–350, Feb. 2019.

[18] N. Wang, M. J. Er, and M. Han "Large tanker motion model identification using generalized ellipsoidal basis function-based fuzzy neural networks," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2732–2743, Dec. 2015.

[19] C. Chen, Z. Liu, K. Xie, Y. Zhang, and C. L. P. Chen, "Asymptotic fuzzy neural network control for pure-feedback stochastic systems based on a semi-Nussbaum function technique," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2448–2459, Sep. 2017.

[20] S. Li and C. Chen, "A regularized monotonic fuzzy support vector machine model for data mining with prior knowledge," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 5, pp. 1713–1727, Oct. 2015.

[21] R. K. Sevakula and N. K. Verma, "Compounding general purpose membership functions for fuzzy support vector machine under noisy environment," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1446–1459, Dec. 2017.

[22] P. Maji and P. Garai, "Fuzzy–rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1166–1177, Aug. 2013.

[23] S. E. Woodward and D. P. Garg, "A numerical optimization approach for tuning fuzzy logic controllers," *IEEE Trans. Cybern.*, vol. 29, no. 4, pp. 565–569, Aug. 1999.

[24] D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 2, pp. 385–389, Apr. 2012.

[25] D. Chen, Q. Hu, and Y. Yang, "Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets," *Inf. Sci.*, vol. 181, no. 23, pp. 5169–5179, 2011.

[26] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, nos. 2–3, pp. 191–209, 1990.

[27] D. Dubois and H. Prade, "Fuzzy sets in approximate reasoning, part 1: Inference with possibility distributions," *Fuzzy Sets Syst.*, vol. 40, no. S1, pp. 143–202, 1991.

[28] W. Wu, J. Mi, and W. Zhang, "Generalized fuzzy rough sets," *Inf. Sci.*, vol. 151, pp. 263–282, May 2003.

[29] F. Xu, D. Miao, and L. Wei, "Fuzzy-rough attribute reduction via mutual information with an application to cancer classification," *Comput. Math. Appl.*, vol. 57, no. 6, pp. 1010–1017, Mar. 2009.

[30] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," *Inf. Sci.*, vol. 502, pp. 18–41, Oct. 2019.

[31] R. R. Yager, "Entropy measures under similarity relations," *Int. J. Gen. Syst.*, vol. 20, no. 4, pp. 341–358, 1992

[32] E. Hernández and J. Recasens, "A reformulation of entropy in the presence of indistinguishability operators," *Fuzzy Sets Syst.*, vol. 128, pp. 185–196, Jun. 2002.

[33] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets," *Knowl. Based Syst.*, vol. 111, pp. 173–179, Nov. 2016.

[34] C. Wang, Q. He, M. Shao, and Q. Hu, "Feature selection based on maximal neighborhood discernibility," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 11, pp. 1929–1940, 2018.

[35] X. Wang, H. Xing, Y. Li, Q. Hua, C. Dong, and W. Pedrycz, "A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 5, pp. 1638–1654, Oct. 2015.

[36] X. Wang, Y. He, and D. Wang, "Non-naive Bayesian classifiers for classification problems with continuous attributes," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 21–39, Jan. 2014.

[37] X. Wang, R. Wang, and C. Xu, "Discovering the relationship between generalization and uncertainty by incorporating complexity of classification," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 703–715, Feb. 2018.

[38] Y. Guo, E. C. C. Tsang, W. Xu, and D. Chen, "Adaptive weighted generalized multi-granulation interval-valued decision-theoretic rough sets," *Knowl. Based Syst.*, vol. 187, Jan. 2020, Art. no. 104804. [Online]. Available: https://doi.org/10.1016/j.knosys.2019.06.012

[39] X. Zhang, C. Mei, D. Chen, and J. Li, "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy," *Pattern Recognit.*, vol. 56, pp. 1–15, Aug. 2016.

[40] B. Bonev, F. Escolano, and M. Cazorla, "Feature selection, mutual information, and the classification of high-dimensional patterns," *Pattern Anal. Appl.*, vol. 11, nos. 3–4, pp. 309–319, 2008.

[41] K. S. Balagani and V. V. Phoha, "On the feature selection criterion based on an approximation of multidimensional mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1342–1343, Jul. 2010.

[42] Q. Hu, L. Zhang, Y. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, Feb. 2018.

[43] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[44] A. Cano, A. Masegosa, and S. Moral. (2005). *ELVIRA Biomedical Data Set Repository*. [Online]. Available: http://leo.ugr.es/elvira/DBCRepository/

[45] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[46] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinformat. Comput. Biol.*, vol. 3, pp. 185–205, Apr. 2005.

[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[48] M. Friedman, "A comparison of alternative tests of significance for the problem of *m* ranking," *Ann. Math. Stat.*, vol. 11, no. 1, pp. 86–92, 1940.

Meng Hu received the B.Sc. and M.Sc. degrees from the Chongqing University of Technology, Chongqing, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Macau University of Science and Technology, Macau, China.

His research interests are focused on feature selection, rough sets, and fuzzy sets.

Eric C. C. Tsang received the B.S. degree in computer studies from the City University of Hong Kong, Hong Kong, in 1990, and the Ph.D. degree in computing from Hong Kong Polytechnic University, Hong Kong, in 1996.

He is an Associate Professor with the Faculty of Information Technology, Macau University of Science and Technology, Macau, China. His current research interests include fuzzy systems, rough sets, fuzzy rough sets, and multiple classifier systems.

Yanting Guo received the B.Sc. degree in mathematics from Jinzhong University, Shanxi, China, in 2014, the M.Sc. degree in applied mathematics from the Chongqing University of Technology, Chongqing, China, in 2017, and the Ph.D. degree in computing from the Macau University of Science and Technology, Macau, China, in 2020.

Her current research interests include information fusion, feature selection, granular computing, decision analysis, and machine learning.

Weihua Xu received the B.Sc. degree in mathematics from Yanbei Normal University, Datong, China, in 2001, the M.Sc. degree in mathematics from Guangxi University, Nanning, China, in 2004, and the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2007.

He is currently a Professor with the College of Artificial Intelligence, Southwest University, Chongqing, China. He has published four academic books and over 120 articles in international journals, and serves on the editorial boards of several international journals. His current research interests include concept cognitive learning, granular computing, approximate reasoning, and uncertainty analysis.