Large-Scale Multimodality Attribute Reduction With Multi-Kernel Fuzzy Rough Sets

Qinghua Hu, *Senior Member, IEEE*, Lingjun Zhang, Yucan Zhou, and Witold Pedrycz, *Fellow, IEEE*

***Abstract*—In complex pattern recognition tasks, objects are typically characterized by means of multimodality attributes, including categorical, numerical, text, image, audio, and even videos. In these cases, data are usually high dimensional, structurally complex, and granular. Those attributes exhibit some redundancy and irrelevant information. The evaluation, selection, and combination of multimodality attributes pose great challenges to traditional classification algorithms. Multikernel learning handles multimodality attributes by using different kernels to extract information coming from different attributes. However, it cannot consider the aspects fuzziness in fuzzy classification. Fuzzy rough sets emerge as a powerful vehicle to handle fuzzy and uncertain attribute reduction. In this paper, we design a framework of multimodality attribute reduction based on multikernel fuzzy rough sets. First, a combination of kernels based on set theory is defined to extract fuzzy similarity for fuzzy classification with multimodality attributes. Then, a model of multikernel fuzzy rough sets is constructed. Finally, we design an efficient attribute reduction algorithm for large scale multimodality fuzzy classification based on the proposed model. Experimental results demonstrate the effectiveness of the proposed model and the corresponding algorithm.**

***Index Terms*—Fuzzy rough sets, multikernel learning, multimodality attribute reduction, parallel computing.**

I. Introduction

IN CURRENT applications, most pattern recognition tasks involve data that are heterogeneous and exhibit multimodality. Those include categorical, numerical, image, text, audio, and even video information. In the era of big data, it is widely accepted that more than 80% of information is carried by heterogeneous and unstructured data. For example, in medical diagnosis systems, there exit categorical attributes of the examination index, numerical attributes of blood pressure, time series of electrocardiography (ECG), and images of B ultrasonic and Computed Tomography (CT) imaging. It becomes a challenging task to evaluate, select, and combine these attributes [1]. Although doctors easily exploit such information, existing machine learning algorithms cannot process it effectively. By 2020, data content will comprise a mixture of text, speech, still and video images, histories of interactions with friends, information sources and their automated proxies, and tracks of sensor readings from global positioning system devices, medical devices, and other embedded sensors in our environment [2]. Thus, it becomes highly desirable to develop an effective representation model and an evaluation strategy for multimodality pattern recognition tasks.

Attribute concatenation and classifier ensemble are two typical methods for handling multimodality data. The first method concatenates different attributes into a long vector and views them as inputs coming into a single classifier. The second method inputs the attributes into different classifiers and then votes or averages the outputs of these classifiers are formed. Attribute concatenation usually leads to ultrahigh dimensionality and ignores the structural information of data, while ensemble learning is sensitive to the choice of classifiers. Moreover, the interaction information among different attributes is not fully captured.

To avoid these problems, multikernel learning has been proposed to transform heterogeneous attributes into a unified representation framework [3]. Kernel functions are employed to quantify similarity, and then these kernels are combined using a certain fusion strategy. A set of kernel functions are designed to quantify the similarity between samples described by different types of attributes. A match kernel is used to construct an equivalence relation [4]; a string kernel is used to calculate the similarity of two strings in gene analysis [5]; a histogram intersection kernel is used to match images [6]. In recent years, many studies have been reported on multikernel learning [7], attribute reduction [8]–[10], and classification [11], [12]. However, these methods do not consider the fuzziness and inconsistency inherently present in multimodality data.

In traditional classification tasks, the labels of samples are Boolean. The samples either completely belongs to one class or another. However, in some complex cases, the samples can be grouped into multiple labels. For example, a facial expression is often associated with multiple emotions, where membership functions of all generic emotions (e.g., happiness, sadness, surprise, fear, anger, and disgust) are used to describe the complex expression. We may not be able to precisely describe a facial

Manuscript received January 29, 2016; revised May 23, 2016 and July 15, 2016; accepted September 29, 2016. Date of publication January 4, 2017; date of current version February 1, 2018. This work was supported by the National Program on Key Basic Research Project under Grant 2013CB329304, National Natural Science Foundation of China under Grant 61222210 and Grant 61432011, U1435212, and the Program for New Century Excellent Talents in University (no. NCET-12-0399).

Q. Hu, L. Zhang, and Y. Zhou are with the Tianjin Key Lab of Cognition Computing and Applications, Tianjin University, China (e-mail: huqinghua@tju.edu.cn; zhanglingjun@tju.edu.cn; zhouyucan@tju.edu.cn).

W. Pedrycz is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB, Canada, the Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University Jeddah, Jeddah 21589, Saudi Arabia, and also with the Systems Research Institute, Polish Academy of Sciences, Warsaw 02-668, Poland (e-mail: wpedrycz@ualberta.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2017.2647966

expression in terms of a single class. The face might exhibit looks angry and disgust, but at different memberships grades. This task comes as form of a multimodality fuzzy classification.

Rough sets provide a sound model for handling inconsistent information in classification [13]. In 1990, Dubois *et al.* proposed the model of fuzzy rough sets [14]. In fuzzy rough sets, we use a family of fuzzy sets to approximate a fuzzy class. The difference between fuzzy upper and lower approximations describes the inconsistency present in fuzzy classification [15]–[18]. In 2011, Hu *et al.* employed kernel functions to extract fuzzy relations and proposed the definitions of kernelized fuzzy rough sets. These models forms a bridge between kernel machines and rough sets [19]. It was reported, however, that fuzzy rough sets are sensitive to noisy samples. To alleviate this shortcoming, robust fuzzy rough sets were developed [20]. However, the proposed models cannot be used to cope with multimodality data.

In practice, most classification tasks come with heterogeneous attributes, including categorical and numerical features. Heterogeneous attribute reduction has attracted considerable attention. In 2007, Hu *et al.* presented a hybrid attribute reduction method based on fuzzy rough sets and information granulation [21], and in 2008, they designed a heterogeneous feature selection algorithm based on neighborhood rough sets (NRS) [22]. In 2014, Chen and Yang proposed an attribute reduction algorithm to handle categorical and numerical data using a discernibility matrix based on a combination of classical and fuzzy rough sets [23]. In 2015, Qian *et al.* proposed a fuzzy granular structure distance that could effectively discriminate between any two fuzzy granular structures [24], and developed an efficient feature selection algorithm for such hybrid data [25]. These methods consider both categorical and numerical data. In practical applications, multimodality of data involves text, image, audio, and even video information, where each attribute is represented by a set of structured features.

In the era of big data, the size of multimodality data is usually very large. It is time-consuming to perform efficient attribute evaluation and reduction. With this regard MapReduce is a popular parallel computing model [26]. In 2010, Yang *et al.* presented an attribute reduction method for massive data based on MapReduce in the context of rough sets [27]. In 2012, Zhang *et al.* proposed a parallel method for computing rough sets approximations [28]. These parallel algorithms scale well and efficiently handle large-scale data. However, in the above studies the designed algorithms are able to deal only with categorical attributes.

From the above analysis one can conclude that large-scale multimodality attribute reduction for fuzzy classification suffers from several essential shortcomings: 1) multikernel learning are not valid for handling fuzzy classification tasks, and 2) no parallel algorithm for large-scale attribute reduction exists.

To illustrate the proposed framework one can refer to Fig. 1. In a medical system, there are patients characterized by P multimodality attributes. The values of the attributes may be vectors, matrices, time series, images, or protein structures. Then each base kernel k_i determines the similarity matrix between samples with respect to the corresponding attribute. The combination kernel K then integrates the P similarity functions into a unified matrix, which is used to compute multikernel fuzzy rough sets. The proposed model will reduce to the "classical" fuzzy rough sets if all the attribute are numerical.

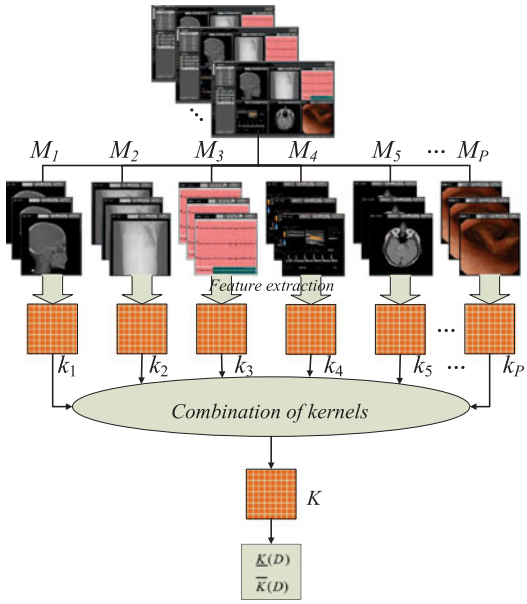


Fig. 1. Framework of multikernel fuzzy rough sets.

In this study, we build a framework for large-scale multimodality attribute reduction with multikernel fuzzy rough sets. The objective is to develop an algorithm to select the informative modalities in multimodality classification. Thus, we evaluate and select a subset of modalities, where each modality may be a vector, a matrix, or an image. This situation is different from the one encountered in the traditional feature selection algorithms. The proposed algorithm reduces to traditional algorithms if the attributes are numerical. The contributions of this paper are twofold. First, using logic operators, we develop a novel combination of kernels techniques and propose a model of multikernel fuzzy rough sets. Second, we describe a parallel strategy to handle large-scale multimodality fuzzy data attribute reduction based on the proposed model.

The paper is organized as follows: In Section II, we present some preliminaries on multikernel learning and fuzzy rough sets. In Section III, we define the combination of kernels and propose the multikernel fuzzy rough set model. In Section IV, the design of the multimodality attribute reduction algorithm for fuzzy classification is described. The experiments are presented in Section V. Finally, conclusions and future work are given in Section VI.

II. Preliminaries

In this section, we introduce multikernel learning, and then review some definitions of rough sets, fuzzy rough sets, and kernelized fuzzy rough sets.

A. Multikernel Learning

Multikernel learning was proposed for handling multimodality attributes by combining multiple kernels. Each kernel is used

TABLE I
T-NORMS AND t-CONORMS

	T-norm	t-conorm
1	$T_M(a,b) = \min(a,b)$	$S_M(a,b) = \max(a,b)$
2	$T_P(a,b) = a \times b$	$S_P(a,b) = a + b - ab$
3	$T_L(a,b) = \max(a+b-1,0)$	$S_L(a,b) = \min(a+b,1)$
4	$T_{\cos}(a,b) = \max(ab - \sqrt{1-a^2}\sqrt{1-b^2}, 0)$	$S_{\cos}(a,b) = \min(a+b-ab+\sqrt{2a-a^2}\sqrt{2b-b^2}, 1)$

to extract information from a certain type of attribute [29]. The fusion of attributes can be interpreted as the combination of base kernels. Given a training set $\{x_i, y_i\}_{i=1}^N$, where y_i is the label of sample x_i for a test sample x, the discriminant function of multikernel learning is computed as

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K_\beta(x_i, x) + b\right)$$

where α is the vector of dual variables corresponding to each separation constraint, b is the bias term for separating the hyperplane, and $K_\beta(x_i, x)$ is computed as

$$K_\beta(x_i, x) = f_\beta(\{k_m(x_i^m, x^m)\}_{m=1}^P)$$

where the combination function $f_\beta : \Re^P \to \Re$ is a linear or nonlinear function to combine P base kernels induced from P sets of attributes. Kernel function $\{k_m : \Re^{D_m} \times \Re^{D_m} \to \Re\}_{m=1}^P$ takes P attributes of sample $\{x^m : \Re^{D_m}\}_{m=1}^P$, where D_m is the dimensionality of the corresponding attribute.

1. Linear combination

$$K_\beta(x_i, x_j) = f_\beta(\{k_m(x_i^m, x_j^m)\}_{m=1}^P) = \sum_{m=1}^{P} \beta_m k_m(x_i^m, x_j^m)$$

where β_m denotes the weights of k_m. The combination methods can be the linear sum ($\beta_m \in \Re$), the conic sum ($\beta_m \geq 0 \in \Re$), or the convex sum ($\beta_m \geq 0 \in \Re$, and $\sum_{m=1}^P \beta_m = 1$). The weights of kernels can also be constrained by the l_p-norm or a trace, for example, the l_1-norm results in a sparse solution, which can be considered as an embedded feature selection algorithm.

2. Nonlinear combination methods

$$K_\beta(x_i, x_j) = f_\beta(\{k_m(x_i^m, x_j^m)\}_{m=1}^P) = \prod_{m=1}^{P} \beta_m k_m(x_i^m, x_j^m)$$

where β_m denotes the weights of k_m.

Multikernel learning is an effective multimodality data learning framework. Multimodality data also contain various inconsistencies, fuzziness, and uncertainties. However, no technique for these problems has been proposed.

B. Fuzzy Rough Sets

Rough sets are considered a powerful model for handling inconsistent information. IS $= \langle U, C\rangle$ is an information system, where U is a sample set and C is a set of condition attributes. U is partitioned into a family of equivalence classes $[x]_R$ by an equivalence relation R derived with the attributes.

Given $x, y, z \in U$, equivalence relation R satisfies the following conditions $R(x,x) = 1$, $R(x,y) = R(y,x)$, and $R(x,y) = 1, R(y,z) = 1 \Rightarrow R(x,z) = 1$.

Given $X \subseteq U$, the lower and upper approximations of X are defined as

$$\underline{R}X = \{[x]_R | [x]_R \subseteq X\}$$
$$\overline{R}X = \{[x]_R | [x]_R \cap X \neq \emptyset\}$$

where $\underline{R}X$ is also called the positive region of X and $\text{BND}_R X = \overline{R}X - \underline{R}X$ is called the boundary of X.

In Pawlak rough sets, condition attributes are discrete, which generate equivalence relations over U. However, most classification tasks are described with numerical or fuzzy data, which cannot be processed directly by Pawlak rough sets. Fuzzy rough sets are proposed based on fuzzy relations.

Given $a, b, c \in [0,1]$, an operator $T : [0,1]^2 \to [0,1]$ is called the triangular norm (T-norm), if it is increasing, associative, commutative, and satisfies $T(a,1) = a$. An operator $S : [0,1]^2 \to [0,1]$ is called a triangular conorm (t-conorm) if it satisfies the first three conditions and $S(a,0) = a$. An operator N (negation) is decreasing and satisfies $N(0) = 1, N(1) = 0$. Some common T-norm and t-conorm operators are listed in Table I.

Given $x, y, z \in U$, the fuzzy $T-$equivalence relation R satisfies the conditions $R(x,x) = 1$, $R(x,y) = R(y,x)$, and $T(R(x,y), R(y,z)) \leq R(x,z)$.

We first give the most general form of fuzzy rough sets. Given a fuzzy approximation space IS $= \langle U, R\rangle$, X is a fuzzy set on U, the lower and upper fuzzy approximations of X are defined as [17]

$$\underline{R_S}X(x) = \inf_{y\in U} S(N(R(x,y)), X(y))$$
$$\overline{R_T}X(x) = \sup_{y\in U} T(R(x,y), X(y)).$$

There are two special cases worth considering. One is that approximate computing is used to approximate fuzzy approximation space in clear conditional attributes space.

Given an approximation space IS $= \langle U, R\rangle$, X is a fuzzy set on U, R is a Boolean relation over U. The lower and upper fuzzy approximations are defined as [17]

$$\underline{R_S}X(x) = \inf_{y\in[x]} X(y)$$
$$\overline{R_T}X(x) = \sup_{y\in[x]} X(y).$$

The other one is that approximate computing is used to approximate clear approximation space in fuzzy conditional attributes space.

Given a fuzzy approximation space IS $= \langle U, R\rangle$, X is a crisp subset of U, R is a fuzzy relation. The lower and upper fuzzy approximations are defined as [17]

$$\underline{R_S}X(x) = \inf_{y\notin X} N(R(x,y)$$
$$\overline{R_T}X(x) = \sup_{y\notin X} R(x,y).$$

Given a nonempty and finite set U, a real-valued function $k : U^2 \to \Re$ is said to be a kernel if k is symmetric and semipositive definite. Moser [30] showed that any kernel function $k : U^2 \to [0,1]$ with $k(x,x) = 1$ is at least $T_{\cos}$-transitive. Hence, fuzzy T-equivalence relation R can be substituted by the relation computed with a kernel function k, which satisfies the following conditions: $k(x,x) = 1$, $k(x,y) = k(y,x)$ and $T_{\cos}(k(x,y), k(y,z)) \le k(x,z)$. This theorem establishes a linkages between kernel machines and fuzzy rough sets.

The kernel functions that satisfy the above conditions can be used to extract the fuzzy $T_{\cos}$-equivalence relations between samples [19]. In fact, a collection of kernel functions, such as the Gaussian kernel, exponential kernel, Laplacian kernel, ANOVA kernel of radial basis function kernel, rational quadratic kernel, circular kernel, and spherical kernel, satisfy these conditions [31].

Given a fuzzy set X, kernelized fuzzy lower and upper approximations are defined as [19]

$$\underline{k_S}X(x) = \inf_{y\in U} S(N(k(x,y)), X(y))$$
$$\overline{k_T}X(x) = \sup_{y\in U} T(k(x,y), X(y)).$$

III. Multikernel Fuzzy Rough Sets

In this section, we translate the idea of multikernel learning to fuzzy rough sets. First, some common kernel functions for multimodality attributes are introduced. Then, the combination of kernels based on fuzzy operator is defined and a model of multikernel fuzzy rough sets is proposed.

A. Kernel Functions for Multimodality Attributes

Categorical, numerical, image, text, and audio are common data modalities. The basic attribute and the applied kernel functions of these multimodality attributes are listed as follows.

1) Given a pair of samples x and y, a match kernel [4] is used to extract equivalence relations from categorical data
$$k(x,y) = \begin{cases} 0, & \text{if } x \neq y; \\ 1, & \text{if } x = y. \end{cases}$$
2) Gaussian kernel [32] is used for extracting information from numerical data
$$k(x,y) = \exp\left(-\frac{||x-y||^2}{\sigma^2}\right)$$
where σ is the parameter of Gaussian kernel.
3) Histogram intersection kernel [6] is designed for computing the similarities between histograms of image data
$$k(x,y) = \sum_{i=1}^{P} \min(x^i, y^i)$$
where x^i and y^i represent the number of pixels that have colors in the i-th fixed list of P color ranges, $\sum_{i=1}^{P} x^i = 1$, $\sum_{i=1}^{P} y^i = 1$. A color histogram represents the distribution of colors in an image.
4) Cosine kernel [33] is used in term frequency-inverse document frequency (TF-IDF) attributes of texts
$$k(x,y) = \frac{xy^T}{||x||||y||}$$
where x and y are TF-IDF vectors that are composed of the product of a term frequency and the inverse document frequency for each token that appears in a string of the database.
5) Cauchy kernel [34] is used in mel-frequency cepstral coefficients (MFCCs) attributes of audio data
$$k(x,y) = \frac{1}{1+\frac{||x-y||^2}{\sigma}}$$
where x and y are the cepstral coefficients.

It is easy to note that these kernels satisfy $k(x,x) = 1$ and $k(x,y) = k(y,x)$, and kernel functions $k : U^2 \to [0,1]$. Therefore, these kernel functions are reflexive, symmetrical, and $T_{\cos}$-transitive according to the theory in [30]. Thus, the fuzzy relations computed with them are fuzzy $T_{\cos}$-transitive relations.

B. Combination of Kernels

In rough sets, the relations derived from two attributes are computed by using the intersection operation. In fuzzy rough sets, the intersection operation corresponds to a T-norm operation. The min operation is a special case of T-norm. We propose a combination of kernels based on T-norm.

MIS $= \langle U, \text{MC}\rangle$ is a multimodality information system and MC $= \{M_1, M_2, ..., M_P\}$ is a set of multimodality condition attributes containing P different attributes; the dimensionality of each attribute may be different.

Given a multimodality information system MIS $= \langle U, \text{MC}\rangle$, kernel k_i is computed for attribute M_i, $i = 1, 2, \ldots, P$. The matrix of kernel function k_i is represented as

$$\begin{bmatrix} k_i(x_1,x_1) & k_i(x_1,x_2) & \cdots & k_i(x_1,x_N) \\ k_i(x_2,x_1) & k_i(x_2,x_2) & \cdots & k_i(x_2,x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k_i(x_N,x_1) & k_i(x_N,x_2) & \cdots & k_i(x_N,x_N) \end{bmatrix}$$

where k_i is computed by a kernel function, such as the Gaussian kernel $k_1(x_i,x_j) = \exp(-\frac{||x_i-x_j||^2}{\sigma^2})$.

Definition 1: Given two kernels k_i and k_j computed with attributes M_i and M_j. For samples $x, y \in U$ four kinds of combination based on the fuzzy T-norm operations are defined as

1) Min
$$K_{T_m}(x,y) = \min(k_i(x,y), k_j(x,y)). \quad (1)$$
2) Product
$$K_{T_p}(x,y) = k_i(x,y) \times k_j(x,y). \quad (2)$$
3) Lukasiewicz T-norm
$$K_{T_l}(x,y) = \max(k_i(x,y) + k_j(x,y) - 1, 0). \quad (3)$$

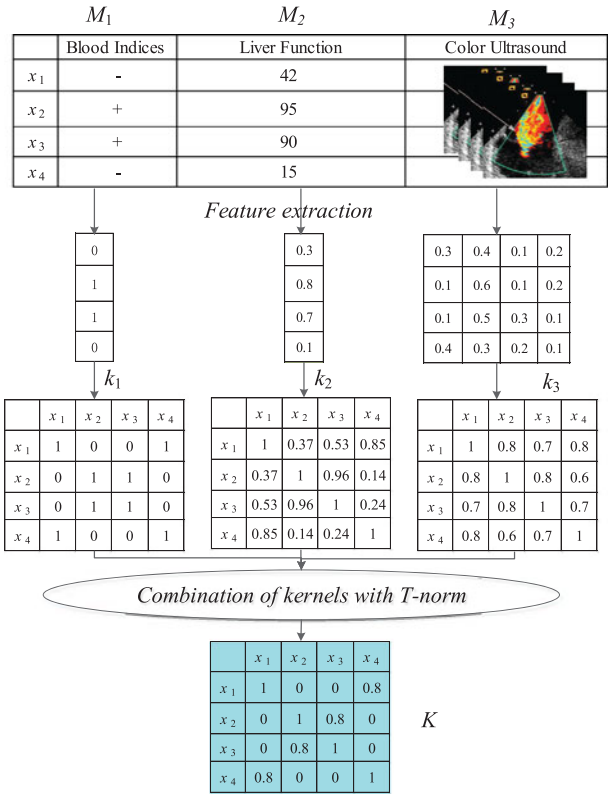


Fig. 2. Combination of kernels.

4) $T_{\cos}$-norm

$$K_{T_{\cos}}(x,y) = \max(k_i(x,y) * k_j(x,y) - \sqrt{1-k_i(x,y)^2}\sqrt{1-k_j(x,y)^2}, 0). \quad (4)$$

Theorem 1: Given $k_i, k_j \in [0,1]$ for M_i and M_j that satisfy $k_i(x,x) = 1, k_j(x,x) = 1$, we have

$$\begin{aligned} K_{T_m} &\subseteq k_i, K_{T_m} \subseteq k_j \\ K_{T_p} &\subseteq k_i, K_{T_p} \subseteq k_j \\ K_{T_l} &\subseteq k_i, K_{T_l} \subseteq k_j \\ K_{T_{\cos}} &\subseteq k_i, K_{T_{\cos}} \subseteq k_j. \end{aligned} \quad (5)$$

Proof: We just take K_{T_m} as an example. Assume that $\text{MB} = M_i \bigcup M_j$. $K_{T_m}^{\text{MB}} = \min\{k_i, k_j\}$. Therefore, $k_i \geq K_{T_m}^{\text{MB}}$ and $k_j \geq K_{T_m}^{\text{MB}}$. That is, $K_{T_m} \subseteq k_i, K_{T_m} \subseteq k_j$. We can prove the other properties in the same manner.

Fig. 2 shows an example of multimodality classification. Four patients are described using the three attributes M_1, M_2, and M_3, where M_1 is blood biochemical indices expressed by categorical attribute, M_2 is liver function indices expressed by a numerical attribute, and M_3 is the color ultrasound images. Thus, we consider match kernel k_1, Gaussian kernel k_2, and histogram intersection kernel k_3 to compute the fuzzy relation between the samples, as an example, consider x_1 and x_2

$$\begin{aligned} k_1(x_1,x_2) &= 0 \\ k_2(x_1,x_2) &= \exp\left(-\frac{||0.3-0.8||^2}{0.5^2}\right) = 0.37 \\ k_3(x_1,x_2) &= \min(0.3,0.1) + \min(0.4,0.6) \\ &\quad + \min(0.1,0.1) + \min(0.2,0.2) = 0.8. \end{aligned}$$

We combine the three attributes with T_m, thus, we get

$$K_{T_m}(x_1,x_2) = \min(0, 0.37, 0.8) = 0.$$

C. Multikernel Fuzzy Rough Sets

Given a multimodality information system $\text{MIS} = \langle U, \text{MC}\rangle$, $\text{MB} = M_1 \cup M_2$, $k_1, k_2 \in [0,1]$ are kernel functions derived with M_1 and M_2, respectively. The fuzzy relation computed by the combination kernels $K_T^{\text{MB}} = f_T(k_1, k_2)$ is a fuzzy T-equivalence relation, where f_T is a combination function based on T-norm.

Definition 2: Given a multimodality information system $\text{MIS} = \langle U, \text{MC}\rangle$, $\text{MB} \subseteq \text{MC}$, the fuzzy lower and upper approximations of X are defined as

$$\begin{aligned} \overline{K_T^{\text{MB}}}X(x) &= \sup_{y\in U} T\left(K_T^{\text{MB}}(x,y), X(y)\right) \\ \underline{K_T^{\text{MB}}}X(x) &= \inf_{y\in U} S\left(N\left(K_T^{\text{MB}}(x,y)\right), X(y)\right). \end{aligned} \quad (6)$$

Theorem 2: $F(U)$ is the family of fuzzy subsets of U. For any fuzzy subset $X_i \in F(U)$, we have the properties

$$\begin{aligned} \underline{K_T^{\text{MB}}}\left(\bigcap_{i\in I} X_i\right) &= \bigcap_{i\in I} \underline{K_T^{\text{MB}}} X_i \\ \overline{K_T^{\text{MB}}}\left(\bigcup_{i\in I} X_i\right) &= \bigcup_{i\in I} \overline{K_T^{\text{MB}}} X_i. \end{aligned} \quad (7)$$

Theorem 3: Given a multimodality information system $\text{MIS} = \langle U, \text{MC}\rangle$ and $\text{MB} \subseteq \text{MC}$, for $\forall X \in F(U)$ the following statements hold

$$\begin{aligned} &\underline{K_T^{\text{MB}}}X \subseteq X \\ &\overline{K_T^{\text{MB}}}X \supseteq X \\ &\underline{K_T^{\text{MB}}}x(y) = \underline{K_T^{\text{MB}}}y(x) \\ &\left(\overline{K_T^{\text{MB}}}\left(U-\{y\}\right)\right)(x) = \left(\overline{K_T^{\text{MB}}}(U-\{x\})\right)(y) \\ &\underline{K_T^{\text{MB}}}\left(\underline{K_T^{\text{MB}}}X\right) = \underline{K_T^{\text{MB}}}X \\ &\overline{K_T^{\text{MB}}}\left(\overline{K_T^{\text{MB}}}X\right) = \overline{K_T^{\text{MB}}}X. \end{aligned} \quad (8)$$

$\text{MDS} = \langle U, \text{MC} \cup D\rangle$ is a multimodality decision system, where D is a decision attribute. In Boolean classification, for $x \in U$, $d_i(x) = \begin{cases} 0, & x \notin d_i; \\ 1, & x \in d_i. \end{cases}$. In fuzzy classification, $d_i(x)$ takes values in $[0,1]$. We also assume $1 \geq d_i(x) \geq 0$, and $\sum_i d_i(x) = 1$.

Definition 3: Given a multimodality decision system $\text{MDS} = \langle U, \text{MC} \cup D\rangle$ and $\text{MB} \subseteq \text{MC}$, the fuzzy lower and

upper approximations of Boolean decision class are defined as

$$\begin{aligned}\underline{K_T^{\mathrm{MB}}}d_i &= \inf_{y\in D-d_i}\left(1-K_T^{\mathrm{MB}}(x,y)\right)\\ \overline{K_T^{\mathrm{MB}}}d_i &= \sup_{y\in d_i}K_T^{\mathrm{MB}}(x,y).\end{aligned}\tag{9}$$

Definition 4: Given MDS $=\langle U, MC\cup D\rangle$ and MB $\subseteq$ MC, the fuzzy lower and upper approximations of a fuzzy decision class are defined as

$$\begin{aligned}\overline{K_T^{\mathrm{MB}}}d_i(x) &= \sup_{y\in U}T\left(K_T^{\mathrm{MB}}(x,y),d_i(y)\right)\\ \underline{K_T^{\mathrm{MB}}}d_i(x) &= \inf_{y\in U}S\left(N(K_T^{\mathrm{MB}}(x,y)),d_i(y)\right).\end{aligned}\tag{10}$$

Theorem 4: Given MDS $=\langle U,\mathrm{MC}\cup D\rangle$ and MB $\subseteq$ MB$'\subseteq$ MC, we have

$$\begin{aligned}\underline{K_T^{\mathrm{MB}}}d_i &\subseteq \underline{K_T^{\mathrm{MB}'}}d_i\\ \overline{K_T^{\mathrm{MB}}}d_i &\supseteq \overline{K_T^{\mathrm{MB}'}}d_i.\end{aligned}\tag{11}$$

Proof: Take K_{T_m} as an example. Assume that MB $=\bigcup_{m=1}^{P}\{M_m\}$ and MB$'=\bigcup_{m=1}^{P'}\{M_m\}$. As MB $\subseteq$ MB$'$, we have $P\le P'$. $K_{T_m}^{\mathrm{MB}}=\min\{k_m\}_{m=1}^{P}=A$ and $K_{T_m}^{\mathrm{MB}'}=\min\{k_m\}_{m=1}^{P'}=\min\{A,k_{P+1},k_{P+2},...,k_{P'}\}$. Therefore, $K_{T_m}^{\mathrm{MB}}\ge K_{T_m}^{\mathrm{MB}'}$. The proof procedure of other methods of the combination kernels based on T-norm are similar. Then, we have $K_T\supseteq K_T'$, that is, $\underline{K_T^{\mathrm{MB}}}d_i\subseteq\underline{K_T^{\mathrm{MB}'}}d_i$. ■

Definition 5: Given MDS $=\langle U,\mathrm{MC}\cup D\rangle$, D is a decision attribute, and $K_T^{\mathrm{MB}}d_i(x)$ is the membership of x to fuzzy set $K_T^{MB}d_i$, we have $0\le\sum_{i=1}^{l}K_T^{\mathrm{MB}}d_i(x)\le 1$. The membership vector of x to decision class d_i in terms of MB can be expressed as

$$\mathrm{POS}_T^{\mathrm{MB}}(D)(x)=\langle\underline{K_T^{\mathrm{MB}}}d_1(x),\ \underline{K_T^{\mathrm{MB}}}d_2(x),\ldots,\ \underline{K_T^{\mathrm{MB}}}d_l(x)\rangle.\tag{12}$$

Definition 6: Given MDS $=\langle U,\mathrm{MC}\cup D\rangle$, the dependency function of D on MB is defined as

$$\gamma_T^{\mathrm{MB}}(D)=\frac{|\mathrm{POS}_T^{\mathrm{MB}}(D)|}{|U|}\tag{13}$$

where $\left|\mathrm{POS}_T^{\mathrm{MB}}(D)\right|=\sum_{x\in U}\sum_{d_i}\inf_{y\in U}S(N(K_T^{\mathrm{MB}}(x,y)),d_i(y))$.

Theorem 5: Given $\langle U,\mathrm{MC}\cup D\rangle$ and MB $\subseteq$ MB$'\subseteq$ MC, we have

$$\begin{aligned}\mathrm{POS}_T^{\mathrm{MB}}(D) &\subseteq \mathrm{POS}_T^{\mathrm{MB}'}(D)\\ \gamma_T^{\mathrm{MB}}(D) &\le \gamma_T^{\mathrm{MB}'}(D).\end{aligned}\tag{14}$$

Proof: The proof can be derived from the monotonicity of the lower approximations.

D. Discussion on Multikernel SVM and Multikernel Fuzzy Rough Sets

In training the multi-kernel learning model, one can define the primal formulation of the objective function and solve its dual form. The strong duality of the primal formulation reads as follows

$$\begin{aligned}\max_{\alpha}\ &\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K_\beta(x_i,x_j)\\ \text{s.t. }&\sum_{i=1}^{N}y_i\alpha_i=0\\ &\forall i,\ \ 0\le\alpha_i\le C\end{aligned}\tag{15}$$

where α is dual variable. $K_\beta=\sum_{m=1}^{P}\beta_m k_m$ is the combination of P base kernels k_m. β_m is the weight of the m kernel. If the combination of kernels K_β is replaced by a base kernel k_m, the formulation reduces to a standard Support Vector Machine (SVM).

In [35] and [36], Xu *et al.* proposed the optimization problem for feature selection

$$\begin{aligned}\min_{0\le\beta\le 1}\ &\omega(\beta)\\ \text{s.t. }&\beta^T e=s\end{aligned}\tag{16}$$

where $\omega(\beta)$ is the value of the objective function in (15). The subset of s most informative features is chosen by minimizing $\omega(\beta)$. Because of the min–max problem, the optimal solution of β is not a discrete solution.

In order to develop a valid and sparse solution for β, the min–max problem is transformed into the following problem with the constraint on β [37]

$$\begin{aligned}\min_{\beta}\max_{\alpha}\ &\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\sum_{m=1}^{P}\beta_m k_m(x_i^m,x_j^m)\\ \text{s.t. }&\sum_{i=1}^{N}y_i\alpha_i=0\\ &\forall i,\ \ 0\le\alpha_i\le C\\ &||\beta||_1=1,\ \beta\ge 0.\end{aligned}\tag{17}$$

The frequently used one-step method can not output both the weights of the base kernels and the parameters of the base learner in a single pass. Therefore, for training the problem a two-step method is required, which is an iterative approach. In each iteration, the weights of the base kernels are updated when determining the base learner parameters, and then, the base learner's parameters are updated when determining the weights of the base kernels. These two steps are repeated until convergence has been achieved. The two-step methods consume a considerably greater amount of time to solve optimization problems than the one-step algorithms.

Multikernel learning leads to a sparse solution of weights of kernels at the expense of the increased time cost. To arrive at a smaller number of attributes and reduce the time consumed, the greedy algorithm can be used for attribute reduction. A heuristic knowledge that contains the information of the objective function is used to evaluate and select attributes in the search process.

Given a multimodality decision system MDS $=\langle U,\mathrm{MC}\cup D\rangle$, MB $\subseteq$ MB$'\subseteq$ MC, the dependence function $\gamma_\beta(D)$ based

Algorithm 1: Multimodality Fuzzy Data Attribute Reduction.

Input: $MDS = \langle U, MC \cup D \rangle$: multimodality decision system

Output: Red: Attribute subsets

1: $Red \leftarrow \emptyset$
2: **while** $\gamma_T^{Red}(D) = \gamma_T^{MC}(D)$ **do**
3: find $MB \subseteq MC$ by maximizing $\gamma_T^{MB}(D)$;
4: $\mathrm{Red} \leftarrow \mathrm{Red} \cup MB$;
5: $MC = MC \backslash \mathrm{Red}$;
6: **end while**
7: **return** Red;

on K_β is not monotone. Assume that $\mathrm{MB} = \bigcup_{m=1}^{P} \{M_m\}$ and $\mathrm{MB}' = \bigcup_{m=1}^{P'} \{M_m\}$. As $\mathrm{MB} \subseteq \mathrm{MB}'$, we have $P \leq P'$. $K_\beta^{\mathrm{MB}} = \sum_{m=1}^{P} \beta_m k_m = A$ and $K_\beta^{\mathrm{MB}'} = \sum_{m=1}^{P'} \beta_m k_m = A + \sum_{m=P+1}^{P'} \beta_m k_m$. We have the linear sum ($\beta_m \in \Re$), the conic sum ($\beta_m \geq 0 \in \Re$), and the convex sum ($\beta_m \geq 0 \in \Re$ and $\sum_{m=1}^{P} \beta_m = 1$). In the above three conditions, $K_\beta^{\mathrm{MB}} \geq K_\beta^{\mathrm{MB}'}$ may not be true. Therefore, the dependence function $\gamma_\beta(D)$ is not monotone.

According to the analysis above, we know that the proposed dependence function $\gamma_T(D)$ based on the combination of kernels K_T is monotone. Therefore, the dependency we propose can be used as a heuristic knowledge for designing a greedy algorithm.

IV. Multimodality Attribute Reduction for Large-Scale Fuzzy Classification

We describe the proposed multimodality attribute reduction algorithm based on multikernel fuzzy rough sets and develop its parallel algorithm in this section.

A. Multimodality Attribute Reduction

We propose the attribute reduction algorithm to deal with multimodality fuzzy classification. A kernel function is employed to compute the fuzzy equivalence relation of the corresponding attribute. The kernel functions are then combined with fuzzy operators to compute the fuzzy equivalence relation of the multimodality attribute subsets.

In the forward heuristic search, we start with an empty set of attributes and select one best attribute by maximizing (13). We design a multimodality fuzzy data attribute reduction Algorithm 1.

B. Parallel Algorithm

The above algorithm is not efficient for large-scale task while multimodality data are usually very large. We find that the selection of the best attribute in each round can be computed in parallel. So that a large-scale dataset can be handled. We developed a parallel computing in the framework of MapReduce. The flowchart of the parallel algorithm is outlined in Fig. 3.

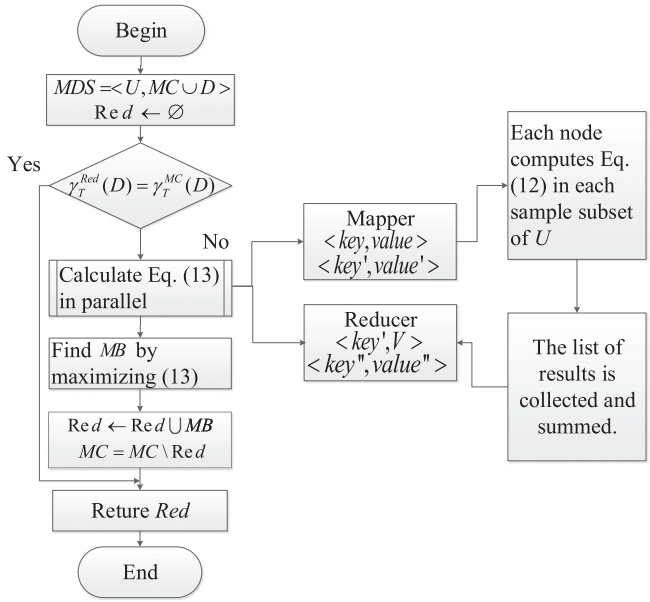


Fig. 3. Parallel attribute reduction algorithm.

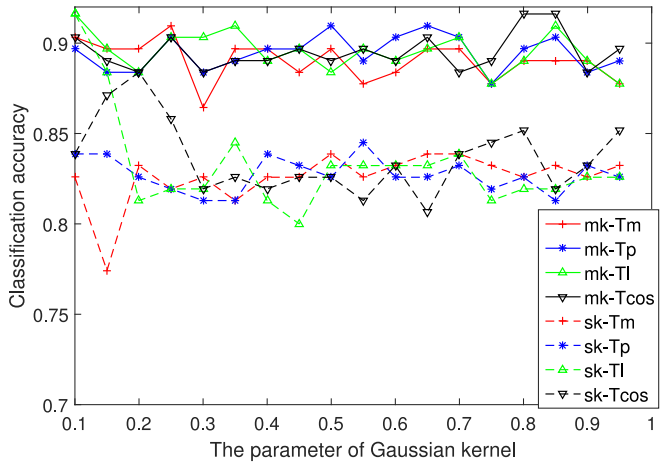


Fig. 4. Classification accuracies of hepatitis versus the values of the parameter of the Gaussian kernel.

Let us recall that MapReduce is a parallel computing model developed by Google for handling large-scale datasets. Map and reduce are the two basic steps in the MapReduce framework [38]. Map and reduce are shown in the proposed parallel algorithm in the form of Algorithms 2 and 3, respectively. In Algorithm 2, each slave node computes (12) for each sample subset of the whole U. In Algorithm 3, the results are collected and summed with the same fixed constant c by Key'. The output of the parallel algorithm is the dependency function of D in terms of MB on U.

V. Experimental Analysis

We report on some experiments to demonstrate the effectiveness of the proposed methods. First, we compare the algorithm with other algorithms considering UCI heterogeneous datasets. Then, we compare the proposed algorithm with some multikernel learning algorithms on multimodality fuzzy classification tasks.

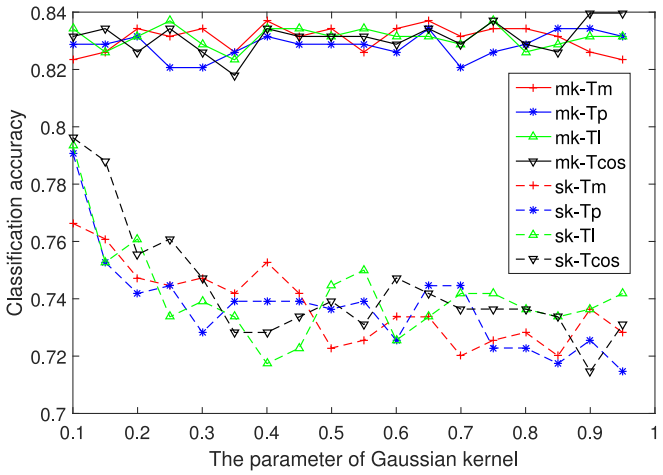


Fig. 5. Classification accuracies of heart versus the values of the parameter of the Gaussian kernel.

Algorithm 2: Map.

Input: key: Attribute subsets Red
$value$: Multimodality decision system MDS
Output: key': a fixed constant c
$value'$: $[R_i, \gamma^{R_i}(D)]$
1: $\left|POS_T^{\mathrm{Red}}(D)(x)\right| = 0$
2: **for** $x \in U' \subseteq U$ **do**
3: Computing $\left|POS_T^{\mathrm{Red}}(D)(x)\right|$ by Eq. (12)
4: **end for**

Algorithm 3: Reduce.

Input: key': a fixed constant c
V: the list of $\left|\mathrm{POS}_T^{\mathrm{Red}}(D)(x)\right|$ from different hosts
Output: key'': a fixed constant c
$value''$: γ
1: Summing $\left|POS_T^{\mathrm{Red}}(D)(x)\right|$ from different hosts;
2: $\gamma = \left|POS_T^{\mathrm{Red}}(D)\right| / |U|$;

It is notable that as to UCI datasets, each feature is considered as an attribute. we compute the kernels with single features and select some of the features. However, as to multimodality datasets, we compute kernel matrices with all the features in a modality and regard the set of features in this modality as a single attribute. In attribute reduction, we select an attribute each round, which means we select a modality each time.

A. UCI Datasets

Some UCI datasets [39] contain both categorical and numerical attributes so that they can be conveniently considered of multimodal nature. We evaluated the performance of the attribute reduction on five UCI datasets with heterogeneous attributes. These datasets are listed in Table II.

In order to analyze the influence of the kernel parameter on classification performance, we experimented with the parameter of the Gaussian kernel located in the interval in [0.1, 0.95] with the step of 0.05. We used two methods to choose the kernel function and ran the proposed attribute reduction. One is multikernel learning, which employs a match kernel function for categorical attributes and a Gaussian kernel function for numerical attributes. The second is a single-kernel method, which just employs the Gaussian kernel function. Four kinds of multikernel combinations, mk-T_m, mk-T_p, mk-T_l, and mk-$T_{\cos}$ are compared with four single-kernel methods, sk-T_m, sk-T_p, sk-T_l, and sk-$T_{\cos}$. The results of two datasets are shown in Figs. 4 and 5, respectively.

TABLE II
DESCRIPTION AND CLASSIFICATION ACCURACY OF FIVE HETEROGENEOUS UCI DATASETS

Data	N	Categorical	Numerical	C	CART	SVM
Anneal	798	6	32	5	99.89 ± 0.35	99.89 ± 0.35
Credit	690	6	9	2	82.73 ± 14.86	81.44 ± 7.18
Heart	270	6	7	2	74.07 ± 6.30	81.11 ± 7.50
Hepatitis	155	6	13	2	91.00 ± 5.45	83.50 ± 5.35
Horse	368	7	15	2	95.92 ± 2.30	72.30 ± 3.63

The experimental results show that both datasets of the multikernel method produces better performance than the single-kernel method. The difference of classification accuracy has nothing to do with the parameters of the Gaussian kernel and the combination techniques of kernels. In order to maintain consistency of comparison, in the following experiment the parameter of the Gaussian kernel was set to 0.2.

We use CART and RBF-SVM in the OSU-SVM 3.00 software package as the classification algorithms and we computed the classification accuracy with the ten-fold cross validation. To handle heterogeneous attributes, we introduce the match kernel function for categorical attributes and the Gaussian kernel function for numerical attributes. We compare the proposed methods with NRS [22] and fuzzy rough sets (GDS) [19]. The multikernel fuzzy rough set methods with different $T-$norm are denoted by T_m, T_p, T_l, and $T_{\cos}$, respectively. The experimental results are given in Tables III and IV, respectively. The best accuracies are shown in boldface.

The experimental results demonstrate that T_m usually produces higher classification accuracies with fewer attributes. However, this conclusion does not hold for all the classification tasks. We should find the optimal operator among the candidates based on experiment.

B. Large-Scale Datasets

We now test the parallel algorithm with some large-scale datasets. We used four datasets in Table V, where the data sets credit and heart are duplicated several times. We still use the match kernel for categorical attributes and the Gaussian kernel for numerical attributes. We ran the algorithm on a cluster of nine nodes, where one was set as a master node and the remaining nodes are configured as slave nodes. Each node had Intel (R) Core (TM) i5-3470 3.2 GHz CPU and 4 GB of main memory and the nodes were connected via an Ethernet network. The size of slave nodes (1–3) was 500 GB, and of slave nodes (4–8) was 1 TB, the total capacity of the cluster was 6 TB.

TABLE III
CLASSIFICATION ACCURACIES BY CART(%) AND THE NUMBER OF SELECTED ATTRIBUTES

Data	NRS	GDS	T_m	T_p	T_l	$T_{\cos}$
Anneal	100.00 ± 0.0(3)	100 ± 0.0(3)	**100.00 ± 0.0**(3)	99.89 ± 0.35(3)	99.89 ± 0.35(3)	99.89 ± 0.35(3)
Credit	82.03 ± 13.5(6)	82.28 ± 14.79(6)	**84.72 ± 12.39**(6)	83.59 ± 11.21(5)	83.59 ± 11.21(5)	83.96 ± 13.68(3)
Heart	75.93 ± 7.66(8)	77.41 ± 8.81(5)	74.98 ± 6.02(6)	**79.63 ± 8.61**(6)	78.89 ± 6.54(6)	76.52 ± 4.54(5)
Hepatitis	90.33 ± 4.57(5)	92.33 ± 6.68(5)	**94.45 ± 5.36**(3)	87.34 ± 6.24(2)	90.86 ± 7.68(3)	85.58 ± 4.91(2)
Horse	95.13 ± 3.96(7)	96.47 ± 1.30(4)	**96.47 ± 1.30**(4)	92.44 ± 6.71(1)	93.88 ± 4.52(2)	92.45 ± 3.47(1)

TABLE IV
CLASSIFICATION ACCURACIES BY RBF-SVM(%) AND THE NUMBER OF SELECTED ATTRIBUTES

Data	NRS	GDS	T_m	T_p	T_l	$T_{\cos}$
Anneal	100.00 ± 0.0(3)	99.89 ± 0.35(3)	**100.00 ± 0.0**(3)	99.89 ± 0.35(3)	99.89 ± 0.35(3)	99.89 ± 0.35(3)
Credit	85.48 ± 18.5(6)	85.92 ± 18.39(5)	**85.94 ± 12.39**(5)	85.77 ± 18.58(3)	85.77 ± 18.58(3)	85.48 ± 18.51(1)
Heart	83.33 ± 6.59(12)	**85.93 ± 6.25**(6)	80.74 ± 8.15(4)	83.33 ± 5.01(4)	78.89 ± 6.54(4)	80.00 ± 7.03(4)
Hepatitis	89.00 ± 4.46(5)	91.67 ± 6.89(3)	**92.33 ± 7.38**(3)	84.50 ± 6.29(1)	89.17 ± 6.54(2)	83.33 ± 3.51(3)
Horse	87.24 ± 3.61(7)	**91.05 ± 3.96**(5)	90.76 ± 5.71(5)	89.11 ± 4.45(1)	81.84 ± 5.12(3)	89.11 ± 4.45(1)

TABLE V
DESCRIPTION OF THE LARGE-SCALE DATASETS AND COMPUTATION TIME OF OUR PROPOSED PARALLEL ALGORITHMS

Data	Instances	Attributes	Classes	Size	Computational time for different numbers of nodes used (S)							
					1	2	3	4	5	6	7	8
Credit	689986	15	2	45 M	47144	27516	22347	18216	16778	15092	14954	13644
Heart	269994	13	2	16 M	19352	13041	11631	10141	9769	9979	10310	9737
KDD	4898431	41	23	682 M	337660	178304	125037	101416	87915	74489	66030	64265
Poker	1000000	38	10	23 M	19443	12343	10363	9052	8926	7874	7749	7689

In this study, map was used to calculate the similarity between a sample and its different-class samples, and therefore, all data files need to be read. If we adopted the hadoop distributed file system (HDFS) [40] to read files directly, reading speed would be very slow because the HDFS data file needs to be read when computing each sample. Considering the efficiency, we store the data file in a distributed cache [41], and then, for convenience inserted all the different-class samples in the array list. The default size of a file data chunk was 64 M, which is not suitable for small datasets. Therefore, we set the size of a file data chunk to 4 M for these datasets the size is less than 64 M.

To measure speedup, the size of dataset is fixed and increased the number of nodes (computers) in the experiment. The speedup is defined as [42]

$$\text{speedup}(p) = \frac{t_1}{t_p}$$

where p is the number of nodes (computers), t_1 is the execution time on a single node, and t_p is the execution time on p nodes.

The ideal result of a parallel algorithm is a linear speedup: A system with p times the number of computers yields a speedup of p. However, linear speedup is difficult to achieve as the communication cost increases as the number of clusters becomes larger. Table V shows the computational time of our proposed parallel algorithms with different nodes. As the number of nodes increases, the computational time of the parallel algorithms becomes shorter. The experimental results shown in Fig. 6 indicate that the speedup improves as the size of the dataset increases. Therefore, the proposed parallel algorithms can treat large-scale data efficiently.

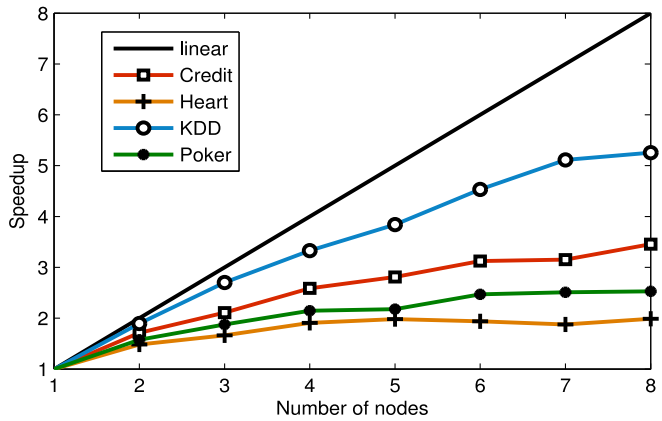


Fig. 6. Speedup of the parallel system.

C. Multimodality Datasets

In this section, we compare the proposed methods with some multikernel learning (MKL) algorithms on two multimodality datasets: the Protein Fold Prediction dataset and NUS-WIDE-

Fig. 7. Example images in Protein Fold Prediction dataset.

Fig. 8. Example images in NUS-WIDE-Object dataset.

TABLE VI
DESCRIPTION OF MULTIMODALITY PROTEIN FOLD PREDICTION DATASET

	Attribute	Data Source	Dimension
1	COM	Amino-acid composition	20
2	SEC	Predicted secondary structure	21
3	HYD	Hydrophobicity	21
4	VOL	Van der Waals volume	21
5	POL	Polarity	21
6	PLZ	Polarizability	21
7	L1	Pseudo amino-acid composition at interval 1	22
8	L4	Pseudo amino-acid composition at interval 4	28
9	L14	Pseudo amino-acid composition at interval 14	48
10	L30	Pseudo amino-acid composition at interval 30	80
11	BLO	Smith-Waterman scores with the BLOSUM 62 matrix	311
12	PAM	Smith-Waterman scores with the PAM 50 matrix	311

Object dataset. Figs. 7 and 8 present some samples coming from the two datasets. It is remarkable that multimodality attributes do not just refer to the raw images, texts, audio, or video. Sometimes they also mean multiple descriptors extracted from the raw data, such as Histogram of Oriented Gradient (HOG), Scale-invariant feature transform (SIFT) and Speed-up robust features (SURF) [8].

The Protein Fold Prediction dataset [3] contains 694 samples belonging to two classes. The samples are described with 12 multimodality attributes. The information of these attributes is summarized in Table VI.

The NUS-WIDE-Object dataset [43] contains 30 000 images of 31 classes. We use five classes *dog*, *fish*, *leaf*, *tower*, and *toy* containing a total of 5 575 images in the experiment. The

TABLE VII
DESCRIPTION OF MULTIMODALITY NUS-WIDE-OBJECT DATASET

	Attribute	Data Source	Dimension
1	CH	Color histogram	64
2	CORR	Color auto-correlogram	144
3	EDH	Edge direction histogram	73
4	WT	Wavelet texture	128
5	CM	Block-wise color moments	255
6	Gaussian	Random Gaussian noise	100
7	Uniform	Random noise subject to uniform distribution	100
8	Chi2	Random noise subject to Chi square distribution	100
9	F-dist	Random noise subject to F-distribution	100
10	Beta	Random noise subject to beta-distribution	100
11	CH+N	Some Gaussian noise is added to the original CH	64
12	CORR+N	Some Gaussian noise is added to the original CORR	144

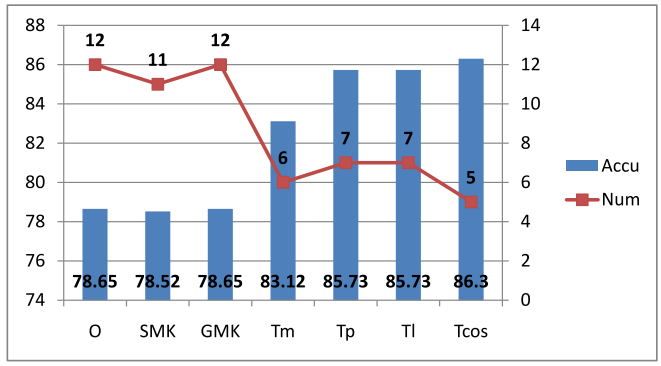


Fig. 9. Classification accuracies by MKL (percent) and the number of attributes of the Protein Fold Prediction dateset.

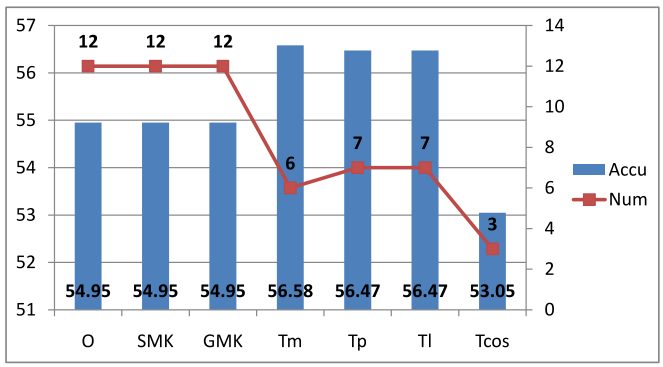


Fig. 10. Classification accuracies by MKL (percent) and the number of attributes of the NUS-WIDE-OBJECT dataset.

samples are described with 12 attributes. The description of these attributes is given in Table VII.

We use MKL [3] as classification algorithm and compute the classification performance with ten-fold cross validation. In the experiment, we try different kernel functions and chose those producing the best performance. As to the Protein Fold Prediction dataset, we try the Gaussian kernel with different parameters, and for the NUS-WIDE-Object dataset, we try the Gaussian kernel with different parameters, histogram intersection kernel, and linear kernel. All the kernel matrices are normalized to unit diagonal.

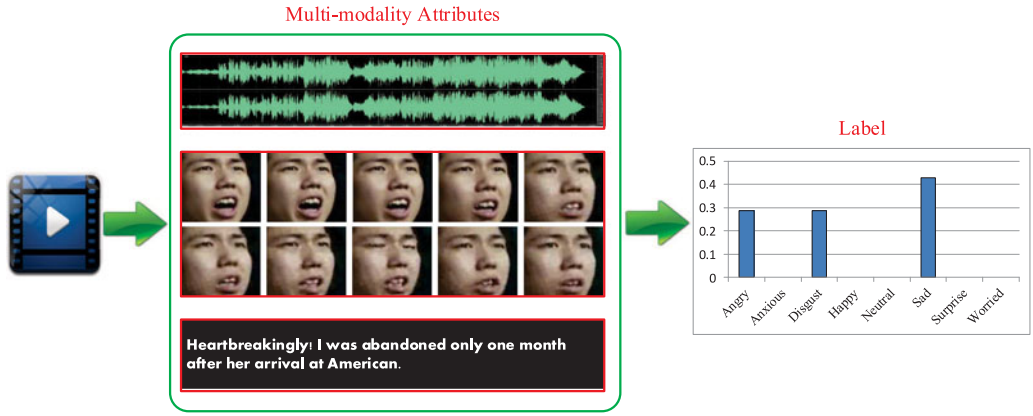


Fig. 11. Example of video with multimodality attributes and fuzzy label.

TABLE VIII
DESCRIPTION OF MULTIMODALITY CHEAVD DATASET

	Attribute	Data Source	Dimension
1	DSIFT_CM	Dense SIFT covariance matrix	6272
2	DSIFT_GD	Dense SIFT gaussian distribution	6272
3	DSIFT_LS	Dense SIFTE linear subspace	6272
4	Gray_CM	Gray covariance matrix	4096
5	HOG_CM	Histogram of oriented gradients covariance matrix	1764
6	HOG_GD	Histogram of oriented gradients gaussian distribution	1764
7	HOG_LS	Histogram of oriented gradients linear subspace	1764
8	LBP_CM	Local binary pattern covariance matrix	59
9	RMS_sma	Root mean square signal frame energy	12
10	RMS_sma_de	Root mean square signal frame energy 1st order delta coefficient	12
11	MFCC_sma	Mel-frequency Cepstral coefficients	144
12	MFCC_sma_de	Mel-frequency Cepstral coefficients 1st order delta coefficient	144
13	ZCR_sma	Zero crossing rate of time signal	12
14	ZCR_sma_de	Zero crossing rate of time signal 1st order delta coefficient	12
15	VP_sma	Voicing probability	12
16	VP_sma_de	Voicing probability 1st order delta coefficient	12
17	F0_sma	Fundamental frequency	12
18	F0_sma_de	Fundamental frequency 1st order delta coefficient	12

We compute the classification performance with the original data (O) and the reduced data by two attribute reduction methods based on SimpleMKL (SMK) [37] and generalized MKL (GMK) [44]. A two-step strategy was used to obtain the classification accuracy after attribute reduction. First, the kernels with weights less than or equal to 0 are removed. Attribute subsets selected by our algorithms are used to train the MKL classifiers. The results are shown in Figs. 9 and 10, respectively.

The results show that the classification accuracy increases after attribute reduction is completed. The classification accuracy rises from 78% to 86.3% once seven attributes have been removed. For the Protein Fold Prediction dataset, the method $T_{\cos}$ produces the highest classification accuracy and selects the least attribute in all methods. For the NUS-WIDE-Object dataset, the method T_m produces the best performance among all methods, while $T_{\cos}$ perform poorly in the same time this model selects attributes. The reason may be that the number of reduced attributes is too small. The other three methods proposed all improve significantly on the original data and other MKL methods.

D. Multimodality Fuzzy Classification

The CHEAVD dataset is a Chinese natural emotional visual–audio database published by the National Laboratory of Pattern Recognition Institute of Automation at Chinese Academy of Sciences (http://www.chineseldc.org/emotion.html). The corpus contains 141 min spontaneous emotional segments extracted from 238 speakers from some films, TV programs, and talk shows.

This dataset contains 1,981 samples of eight kinds of emotions (happy, angry, surprise, disgust, neutral, worried, anxious, sad). Each sample is described by 18 attributes. The first ten descriptors are audio attributes and another eight are visual attributes. In fact, we can extract a lot of other attributes from the videos. The attributes are described in Table VIII. Each video is associated with eight memberships of different emotions. We recognize the emotions of the video segments according to the audio and image information.

MKL [3] is also used to compute the classification accuracies of the original data and reduced data with five-fold cross

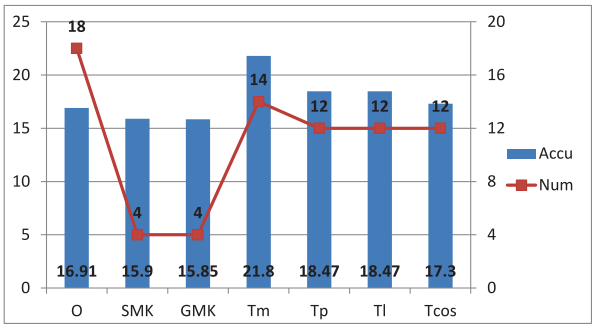


Fig. 12. Classification accuracies by MKL (percent) and the number of attributes of the CHEAVD dateset.

validation. As to audio attributes, the RBF kernel is used to compute the kernel matrices with different parameters. Following [3], the hyperparameter of Gaussian function is set as the number of the features. As to visual attributes, one video clip can be regarded as a set of feature vectors by extracting features form each video frame. Based on the feature vectors, we introduce three types of models: covariance matrix, Gaussian distribution, and linear subspace on visual attributes to describe the video. Finally, Riemannian kernel [45] is used to compute the kernel matrices and the kernel matrices are normalized to unit diagonal.

We compared the proposed method with the original data (O) and the reduced data by two attribute reduction methods based on SMK [37] and GMK [44]. As to SimpleMKL and generalized MKL, we eliminate the attributes if their weights are less than or equal to 0. Then the remaining attribute subsets are used to train the MKL classifier. Fig. 12 gives the classification accuracies and the number of the selected attributes.

From Fig. 12 we conclude that the classification accuracy increases if some irrelevant attributes are removed by the proposed algorithm. The method T_m produces the best performance among all the methods, and it improves the performance from 16.9% to 21.8% when four attributes are removed. However, SimpleMKL just selects four attributes and the accuracy drops from 16.9% to 15.9%, and generalized MKL also just selects four attributes and the accuracy drops to 15.85%.

VI. Conclusion and Future Work

In this study, the model of multikernel fuzzy rough sets was developed by integrating multikernel learning with fuzzy rough sets. We have designed an algorithm of large-scale multimodality attribute reduction based on this model. First, we defined a novel combination of kernels based on T-norm to determine the fuzzy similarity between multimodality data. Then, we have proposed the model of multikernel fuzzy rough sets. Finally, we designed a parallel multimodality attribute reduction algorithm for fuzzy classification. The experimental results show that the proposed algorithm is effective and efficient on large-scale multimodality fuzzy classification tasks.

In the era of big data, objects are usually described with multimodality data, and are associated with complex classification scenarios, such as multilabel classification, fuzzy classification, multigranularity classification, and hierarchical classification [46]–[48]. A promising direction of research is to extend this study to deal with multilabel classification and hierarchical classification. There may be hundreds of labels in some classification task. Furthermore, a set of labels may form a hierarchical structure, and some objects could be associated with multiple labels. For such problems, it becomes challenging to develop efficient algorithms. Moreover, just like in fuzzy emotion recognition, the classification of multimodality data may be inherently fuzzy. The traditional classification algorithms, like SVM and multikernel SVM, cannot deal with this category of problems. It could be interesting to develop fuzzy classification algorithms for such multimodality data by exploiting the proposed multikernel fuzzy rough set.

References

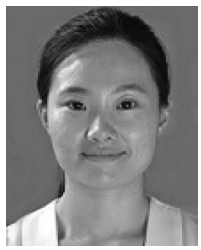
[1] M. F. Balcan, A. Blum, and N. Srebro,"A theory of learning with similarity functions," *Mach. Learn.*, vol. 72, no. 1/2, pp. 89–112, 2008.
[2] P. Norvig *et al.*, "2020 visions," *Nature*, vol. 463, no. 7, pp. 26–32, 2010.
[3] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, 2011.
[4] F. Odone, A. Barla, and A. Verri, "Building kernels from binary strings for image matching," *IEEE Trans. Image Process.*, vol. 14, no. 2, pp. 169–180, Feb. 2005.
[5] H. Lodhi, C. Saunders, J. Shawe Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *J. Mach. Learn. Res.*, vol. 2, pp. 419–444, 2002.
[6] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[7] Y. R. Yeh, T. C. Lin, Y. Y. Chung, and Y. C. F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 563–574, Jun. 2012.
[8] Y. Y. Lin, T. L. Liu, and C. S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.
[9] J. J. Y. Wang, H. Bensmail, and X. Gao, "Feature selection and multikernel learning for sparse representation on a manifold," *Neural Netw.*, vol. 51, pp. 9–16, 2014.
[10] M. M. Liu, W. Sun, and B. Liu, "Multiple kernel dimensionality reduction via spectral regression and trace ratio maximization," *Knowl.-Based Syst.*, vol. 83, pp. 159–169, 2015.
[11] I. M. de Diego, A. Muñoz, and J. M. Moguerza, "Methods for the combination of kernel matrices within a support vector framework," *Mach. Learn.*, vol. 78, no. 1/2, pp. 137–174, 2010.
[12] C. Cortes, M. Mohri, and A. Rostamizadeh, "Multi-class classification with maximum margin multiple kernel," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 46–54.
[13] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.
[14] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2/3, pp. 191–209, 1990.
[15] A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 126, no. 2, pp. 137–155, 2002.
[16] J. S. Mi and W. X. Zhang, "An axiomatic characterization of a fuzzy generalization of rough sets," *Inf. Sci.*, vol. 160, no. 1, pp. 235–249, 2004.
[17] D. S. Yeung, D. G. Chen, E. C. Tsang, J. W. Lee, and X. Z. Wang, "On the generalization of fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 3, pp. 343–361, Jun. 2005.
[18] R. Wang, D. G. Chen, and S. Kwong, "Fuzzy-rough-set-based active learning," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 6, pp. 1699–1704, Dec. 2014.
[19] Q. H. Hu, D. R. Yu, W. Pedrycz, and D. G. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, Nov. 2011.
[20] Q. H. Hu, L. Zhang, S. An, D. Zhang, and D. R. Yu, "On robust fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 4, pp. 636–651, Aug. 2012.

[21] Q. H. Hu, Z. X. Xie, and D. R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognit.*, vol. 40, no. 12, pp. 3509–3521, 2007.
[22] Q. H. Hu, D. R. Yu, J. F. Liu, and C. X. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, 2008.
[23] D. G. Chen and Y. Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 5, pp. 1325–1334, Oct. 2014.
[24] Y. H. Qian, Y. B. Li, J. Y. Liang, G. P. Lin, and C. Y. Dang, "Fuzzy granular structure distance," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 6, pp. 2245–2259, Dec. 2015.
[25] Y. H. Qian, Q. Wang, H. H. Cheng, J. Y. Liang, and C. Y. Dang, "Fuzzy-rough feature selection accelerator," *Fuzzy Sets Syst.*, vol. 258, pp. 61–78, 2015.
[26] W. Z. Zhao, H. F. Ma, and Q. He, "Parallel k-means clustering based on mapreduce," in *Proc. 1st Int. Conf. Cloud Comput.*, 2009, pp. 674–679.
[27] Y. Yang, Z. R. Chen, Z. Liang, and G. Y. Wang, "Attribute reduction for massive data based on rough set theory and mapreduce," in *Proc. Rough Set Knowl. Technol.*, 2010, pp. 672–678.
[28] J. B. Zhang, T. R. Li, D. Ruan, Z. Z. Gao, and C. B. Zhao, "A parallel method for computing rough set approximations," *Inf. Sci.*, vol. 194, pp. 209–223, 2012.
[29] W. S. Noble *et al.*, "Support vector machine applications in computational biology," *Kernel Methods Comput. Biol.*, pp. 71–92, 2004.
[30] B. Moser, "On representing and generating kernels by fuzzy equivalence relations," *J. Mach. Learn. Res.*, vol. 7, pp. 2603–2620, 2006.
[31] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, vol. 2, no. 2, pp. 299–312, 2002.
[32] B. Schlkopf and A. Smola, "Learning with kernels," in *Proc. 21st Int. Conf. Mach. Learn.*, 2001, pp. 639–646.
[33] S. Tata and J. M. Patel, "Estimating the selectivity of TF-IDF based cosine similarity predicates," *ACM Sigmod Record*, vol. 36, no. 2, pp. 7–12, 2007.
[34] L. Lu, B. Huang, Q. Y. Zhang, D. F. Ke, and Y. Y. Xu, "Research on algorithm of combing LDA-based discriminant classifier and MFCC feature extraction for pure acoustic listening similarity," *Int. J. Advancements Comput. Technol.*, vol. 4, no. 5, pp. 106–113, 2012.
[35] Z. L. Xu, R. Jin, J. P. Ye, M. R. Lyu, and I. King, "Non-monotonic feature selection," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1145–1152.
[36] H. Q. Yang, Z. L. Xu, M. R. Lyu, and I. King, "Budget constrained non-monotonic feature selection," *Neural Netw.*, vol. 71, pp. 214–224, 2015.
[37] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008.
[38] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[39] UCI machine learning repository, 2005. [Online]. Available: http://archive.ics.uci.edu/ml.
[40] T. White, *Hadoop : The definitive Guide*, vol. 215, CA, USA: Oreilly Media, Inc., 2010, no. 11, pp. 1–4.
[41] D. Povey and J. Harrison, "A distributed internet cache," *Aust. Comput. Sci. Commun.*, vol. 19, pp. 175–184, 1997.
[42] X. W. Xu, J. Jäger, and H. P. Kriegel, "A fast parallel clustering algorithm for large spatial databases," in *High Performance Data Mining*, New York, NY, USA: Springer, 2002, pp. 263–290.
[43] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 951–958.
[44] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1065–1072.
[45] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interaction*, 2014, pp. 3064–3070.
[46] R. Jensen, S. Vluymans, N. M. Parthalin, C. Cornelis, and Y. Saeys, "Semi-supervised fuzzy-rough feature selection," in *Proc. Rough Sets, Fuzzy Sets, Data Mining, Granular Comput.*, 2015, pp. 185–195.
[47] H. Asfoor, R. Srinivasan, G. Vasudevan, and N. Verbiest, "Computing fuzzy rough approximations in large scale information systems," in *Proc. IEEE Int. Conf. Big Data*, 2014, pp. 9–16.
[48] S. Vluymans *et al.*, "Distributed fuzzy rough prototype selection for big data regression," in *Proc. 2015 Annu. Conf. North Amer. Fuzzy Inf. Process. Soc.*, 2015, pp. 153–158.

Qinghua Hu (SM'13) received the B.S., M.S., and Ph.D degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the Department of Computing, Hong Kong Polytechnic University. He is currently a Full Professor and the Vice Dean of the School of Computer Science and Technology, Tianjin University, Tianjin, China. He has authored more than 100 journal and conference papers in the areas of granular computing based machine learning, reasoning with uncertainty, pattern recognition, and fault diagnosis. His current research interests include rough sets, granular computing, and data mining for classification and regression.

Prof. Hu was the Program Committee Co-Chair of the International Conference on Rough Sets and Current Trends in Computing in 2010, the Chinese Rough Set and Soft Computing Society in 2012 and 2014, and the International Conference on Rough Sets and Knowledge Technology, the International Conference on Machine Learning and Cybernetics in 2014, and the General Co-Chair of IJCRS 2015. He is now PC-Co Chairs of CCML 2017 and CCCV 2017.



Lingjun Zhang received the B.S. and M.S. degree in computer science from Henan Normal University, Xinxiang, China, in 2007 and 2012, respectively. She is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Tianjin University, Tianjin, China. Her research interests are focused on rough sets, granular computing, machine learning for feature selection.

Yucan Zhou received the B.S. and M.S. degree in computer science from Tianjin University, Tianjin, China, in 2012 and 2014, respectively. She is currently working toward the Ph.D degree with the School of Computer Science and Technology, Tianjin University, Tianjin, China.

Her research interests include artificial intelligence, deep learning, and long-tail distribution learning.

Witold Pedrycz (F'99) received the M.Sc., Ph.D., and the D.Sc. degrees from the Silesian University of Technology, Gliwice, Poland.

He is a Professor and the Canada Research Chair of Computational Intelligence with the Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia. He is also with the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. He holds an appointment of special professorship with the School of Computer Science, University of Nottingham, Nottingham, U.K. His current research interests include computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data mining, fuzzy control, pattern recognition, knowledge-based neural networks, relational computing, and software engineering. He has published numerous papers in the above areas. He has also authored 15 research monographs covering various aspects of computational intelligence, data mining, and software engineering.

Dr. Pedrycz received the prestigious Norbert Wiener Award from the IEEE Systems, Man, and Cybernetics Society in 2007, the IEEE Canada Computer Engineering Medal 2008, the Cajastur Prize for Soft Computing from the European Centre for Soft Computing for pioneering and multifaceted contributions to granular computing in 2009, the Killam Prize in 2013, and the Fuzzy Pioneer Award 2013 from the IEEE Computational Intelligence Society. He was elected as a Foreign Member of the Polish Academy of Sciences in 2009 and a fellow of the Royal Society of Canada in 2012. He has been a member of numerous program committees of the IEEE conferences in the area of fuzzy sets and neurocomputing.