This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Multimodal Data Fusion Using Non-Sparse Multi-Kernel Learning with Regularized Label Softening

Peihua Wang, Chengyu Qiu, Jiali Wang, Yulong Wang, Jiaxi Tang, Bin Huang, Jian Su, *IEEE Member*, Yuanpeng Zhang *IEEE Member*

***Abstract*— Due to the need of practical application, multiple sensors are often used for data acquisition, so as to realize the multimodal description of the same object. How to effectively fuse multimodal data has become a challenge problem in different scenarios including remote sensing. Non-sparse multi-kernel learning has won many successful applications in multimodal data fusion due to the full utilization of multiple kernels. Most existing models assume that the non-sparse combination of multiple kernels is infinitely close to a strict binary label matrix during the training process. However, this assumption is very strict so that label fitting has very little freedom. To address this issue, in this study, we develop a novel non-sparse multi-kernel model for multimodal data fusion. To be specific, we introduce a label softening strategy to soften the binary label matrix which provides more freedom for label fitting. Additionally, we introduce a regularized term based on manifold learning to anti over fitting problems caused by label softening. Experimental results on one synthetic dataset, several UCI multimodal datasets and one multimodal remoting sensor dataset demonstrate the promising performance of the proposed model.**

***Index Terms*—Semantic-based multimodal fusion, multi-kernel learning, label softening, manifold learning, remote sensing.**

I. Introduction

Today, with the development of advanced sensors, multimodal data is becoming easier to obtain. Multimodal data refers to the data obtained from different fields or views for the same object, and each field or view describing these data is called a modality [1-2]. For example, in the field of remote sensing [45], full-color images and multispectral images are two modalities than can both be used for earth observation. In general, different modalities of the same object tend to contain complementary information about the description of that object [3]. However, different modalities may lead to significant gaps in performance when methods and algorithms are designed and developed separately. Therefore, how to effectively mine and fuse information extracted from different modalities is still a great challenging task.

Multi-kernel learning as one of the commonly used data fusion methods has been successfully applied in many scenarios [4-5]. To mine enough patterns from training samples, it uses a set of predefined kernels and learns an optimal linear or non-linear combination of them. Typically, when multi-kernel learning is used for classification tasks, it aims to learn a transformation matrix that can transform the combination of kernels into a binary label matrix. Previous studies showed that transforming the combination of kernels into a strict binary label matrix often failed to learn a very discriminant transformation matrix [6-7]. This because such label fitting is very strict and has very little freedom. To address this problem, in this study, a label softening strategy is introduced to relax the binary label matrix. With this strategy, the margins between different classes are enlarged so that label fitting becomes freer. That is to say, transforming the combination of kernels into a soft binary label matrix can help learn a discriminant transformation matrix due to the larger class margins, hence improve classification performance. However, more freedom of label fitting may bring overfitting problems. To address overfitting, a regularized term derived from manifold learning is used to control the training process [8]. The basic idea is that samples in the same class should be kept as close as possible when they are transformed into the label space by multi-kernel techniques. Based on label softening strategy and the manifold regularized term, we develop a novel Non-Sparse Multi-Kernel Learning model with Regularized Label Softening (NS-RLS-MKL) for multimodal data fusion. The contributions can be summarized as follows.

- A label softening strategy is introduced to soften the binary label matrix which provides more freedom for label fitting.
- A regularized term based on manifold learning is introduced to solve over fitting problems caused by label softening.

The rest is organized as follows. In Section II, we briefly review several multimodal data fusion methods. In Section III, we give detailed information about the proposed model including the objective function, optimization, and algorithm steps. In Section IV, extensive experiments are conducted to evaluate the performance of the proposed model and in the last section, we conclude the whole study.

This work was supported in part by the National Natural Science Foundation of China under Grants 82072019, by the Natural Science Foundation of Jiangsu Province under Grant BK20201441, by Jiangsu Post-doctoral Research Funding Program under Grants with No. 2020Z020. (*corresponding author: Yuanpeng Zhang, Jian Su*).

Peihua Wang, Chengyu Qiu, Jiali Wang, Yulong Wang, Jiaxi Tang and Bin Huang are with the Department of medical informatics, Nantong University, Nantong, 226001, China.

Jian Su is the School of Computer and Software at the Nanjing University of Information Science and Technology Nanjing University of Information Science and Technology, Nanjing, China.

Yuanpeng Zhang is with the Department of medical informatics, Nantong University, Nantong, 226001, China (e-mail: maxbirdzhang@ntu.edu.cn).

II. Related Work and Background Concepts

Multimodal data fusion methods can be roughly divided into three categories, stage-based fusion, feature-based fusion, and semantic-based fusion. In this study, we focus on semantic-based fusion. Therefore, in following, we will summarize some works related to semantic-based multimodal data fusion.

A. *Semantic-based multimodal fusion*

The semantic-based multimodal fusion is to understand the data meaning of each modal and the relationship between features from different modalities and use the way of human thinking to abstract the semantic meaning of different modalities to complete multi-modal data fusion. Typically, the existing semantic-based multimodal fusion methods can be divided into three categories: co-training based, multi-kernel learning based, and subspace learning based.

(1) Co-training

Co-training methods [9-11] maximize the synergistic degree of the multi-modal data by alternate training. Co-training is one of the earliest strategies to solve the problem of multi-modal data fusion. In co-training, three assumptions are often required: 1) sufficiency: each modal itself has sufficient data to complete the corresponding analysis task; 2) compatibility: the objective function based on multiple modalities of symbiotic characteristics can predict the same class labels with a high probability; 3) conditional independence: given the specific class label, the modalities are conditionally independent. In practice, the restriction of conditional independence is too strong to satisfy, so some corresponding weak restrictions are proposed. Co-training is based on single-modal learning, and is widely used in the semi-supervised learning fields. For example, multi-modal data is used to iteratively learn multiple classifiers, and the obtained classifiers are applied to each other's unlabeled data classification prediction. Typical multi-modal co-training methods include: Co-EM models based on expectation maximization, support vector machine models based on Co-EM [12], the co-training regression model CoREG [13] and so on.

Original co-training methods cannot test the reliability of class labels obtained from each modality, but in practice even very few samples of incorrect labels may greatly deteriorate the performance of the learned model.

To solve this problem, some scholars have proposed a robust co-training strategy, in which Canonical Correlation Analysis (CCA) is incorporated into the co-training process to check the prediction results of unlabeled data [14]. Yu *et al*. proposed an improved co-training method based on Bayesian undirected graph model, which can query <instance, modality> pairs to improve the performance of learning results [15]. Zhao *et al*. integrated K-means clustering and linear discriminant analysis into the co-training process, and the discriminant subspace of another modality is found through the labeled samples from automatic learning of one modality [16].

(2) Multi-kernel learning

Multi-kernel learning is one of the commonly used kernel-based machine learning strategies. It uses a set of predefined kernel functions to learn an optimized linear or nonlinear combination based on kernel function to complete the analysis of specific data tasks. The kernel is a hypothesis based on data, which may be a concept of similarity, a classifier, or a regressor. According to [17], there are two ways of multi-kernel learning: 1) different kernels correspond to different similarity concepts, and the learning function selects the best kernel calculation results or integrates the calculation results of all kernels. This multi-kernel learning uses all modal data to complete the training of each kernel, which is not suitable for multi-modal fusion learning; 2) different kernels are trained by using different modal data, so the integration of all kernel learning results is equivalent to the fusion of all modal information. Existing multi-kernel result integration algorithms can be divided into three categories: linear, nonlinear and data-dependent integration [18-20].

The main reason why multi-kernel learning is used in multi-modal analysis is that different kernels naturally correspond to different modalities in multi-kernel learning, and proper integration of each kernel can improve the performance of learning results. For example, Poria *et al*. applied multi-kernel learning to multi-modal emotion recognition and semantic analysis [21], and obtained better results than single kernel modal fusion by using different kernels for semantic, video and text modal features. In [22], the authors applied multi-kernel learning to face recognition, and proposed a classification learning algorithm based on multi-kernel sparse representation. The algorithm performs sparse coding and dictionary learning in multi-core space at the same time, and obtains the optimal weight of the kernel through possible kernel combination and sparse coefficient calculation.

(3) Subspace learning

Multimodal data fusion algorithms based on subspace learning assume that all modal data can be projected into the same semantic shared subspace, and data mining tasks such as clustering and classification can be completed in the subspace. Usually, the feature dimension of multi-modal shared subspace is smaller than any one of the dimensions of modal data, so the dimension disaster problem of multimodal data can be solved to a certain extent through low dimensional learning of shared subspace. In the existing papers, the earliest multi-modal shared subspace learning algorithm uses Canonical Correlation Analysis (CCA) to maximize the correlation between two modalities, obtains the maximum correlation subspace, and outputs the projection matrix corresponding to each modality [23]. Based on this, Pereira *et al* [24] proposed the corresponding nonlinear improved algorithm, namely multi-modal shared subspace learning algorithm based on kernel CCA (KCCA). The algorithm first maps the data points to the high-dimensional data space through nonlinear transformation, and then uses linear CCA to complete the subspace learning. Both CCA and KCCA are unsupervised learning algorithms. In [25], a method of multi-modal discriminant analysis (FDA) is proposed to find more effective projection matrix by using labeled information and make the modal data more relevant.

Another effective multi-modal subspace learning algorithm is the algorithm based on matrix decomposition [26-27]. In the single modal data analysis, matrix decomposition decomposes the original data into basis matrix and potential feature representation, and can use the matrix to explain the potential elements learned. Therefore, we can complete different data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

learning tasks through each regularly decomposed matrix. With people's attention to the problem pf multimodal data fusion, more and more multimodal data fusion algorithms based on joint matrix factorization have been offered [28-29].In the existing matrix factorization algorithms, Nonnegative Matrix Factorization (NMF) uses the idea of combining parts to form a whole to reconstructs each data record through the linear combination of nonnegative basis matrix, which is in line with the physiological and psychological perception process of human brain. Now it has been widely used in the potential feature learning of data [30]. For example: The multi-modal nonnegative matrix decomposition method MultiNMF is proposed in [31]. It uses the combined nonnegative matrix decomposition learning to obtain the shared characteristics of multimodal data. In addition, the regularization strategy is also introduced in the MultiNMF algorithm to make the results of different modal learning comparable and ensure the consistent cross-modal sharing characteristics. MMNMF(Multi-Manifold NMF) [32], another multimodal fusion algorithm based on NMF, integrates the uniform manifold structure and the uniform correlation matrix to normalize the multi-manifold structure in the process of non-negative matrix decomposition, so as to ensure the local geometric structure of each modal space in the process of multimodal learning, and obtain more accurate modal fusion features. In addition, some scholars have learned the shared subspace of multimodal data based on Gaussian process [33-34], spectral embedding [35] and undirected graph model [43-44] and achieved good results.

B. Background Concepts

Since the proposed model NS-RLS-MKL is derived from kernel ridge regression, in this section, we will give some background concepts.

By introducing the reproducing kernel Hilbert space, the original Ridge regression can be updated into its kernel version, that is [36],

$$\min_{\mathbf{A}} \left\|\mathbf{X}_\phi \mathbf{A}-\mathbf{Y}\right\|_2^2+\lambda\left\|\mathbf{A}\right\|_F^2 \tag{1}$$

where $\mathbf{X}_\phi=[\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),...,\phi(\mathbf{x}_N)]^T\in R^{N\times d_\phi}$ in which $\phi(\mathbf{x}_i)$ is the mapping of $\mathbf{x}_i$ in the reproducing kernel Hilbert space. Like Ridge regression, the solution to kernel Ridge regression can be deduced as

$$\mathbf{A}^*=(\mathbf{X}_\phi{}^T\mathbf{X}_\phi+\eta\mathbf{I}_{d_\phi})^{-1}\mathbf{X}_\phi\mathbf{Y} \tag{2}$$

We know that the dimension of the reproducing kernel Hilbert space is infinite. Therefore, it is usually infeasible to find the mapping function ϕ. Therefore, instead of directly optimizing (1), an alternative method is to optimize its dual problem, that is,

$$\begin{aligned}&\min_{\mathbf{A},\ \xi}\frac{1}{2}\sum_{i=1}^{N}\xi_i^2+\frac{\lambda}{2}\left\|\mathbf{A}\right\|_F^2\\&s.t.\ \phi(\mathbf{x}_i)^T\mathbf{A}=y_i-\xi_i\ ,i=1,2,...,N\end{aligned} \tag{3}$$

where ξ_i is the training error of sample $\mathbf{x}_i$. By introducing Lagrangian multipliers, the corresponding Lagrangian function of the dual problem is

$$\begin{aligned}J(\mathbf{A},\xi_i,\alpha_i)=&\frac{1}{2}\sum_{i=1}^{N}\xi_i^2+\frac{\lambda}{2}\left\|\mathbf{A}\right\|_F^2\\&-\sum_{i=1}^{N}\alpha_i(\phi(\mathbf{x}_i)^T\mathbf{A}-y_i+\xi_i)\end{aligned} \tag{4}$$

where α_i is the Lagrangian multipliers. With the Karush-Kuhn-Tucker (KKT) and by setting $\partial J(\mathbf{A},\xi_i,\alpha_i)/\partial\mathbf{A}=0$, $\partial J(\mathbf{A},\xi_i,\alpha_i)/\partial\xi_i=0$, and $\partial J(\mathbf{A},\xi_i,\alpha_i)/\partial\alpha_i=0$, we have

$$\mathbf{A}=\mathbf{X}_\phi^T\boldsymbol{\alpha}/\lambda \tag{5}$$

$$\boldsymbol{\xi}=\boldsymbol{\alpha} \tag{6}$$

$$\mathbf{X}_\phi\mathbf{A}-\mathbf{Y}+\boldsymbol{\xi} \tag{7}$$

By substituting (5) and (6) into (7), we have

$$\boldsymbol{\alpha}^*=\lambda(\mathbf{X}_\phi\mathbf{X}_\phi{}^T+\lambda\mathbf{I}_N)^{-1}\mathbf{Y} \tag{8}$$

Therefore, by substituting (8) into (5), we have

$$\mathbf{A}^*=\mathbf{X}_\phi{}^T(\mathbf{X}_\phi\mathbf{X}_\phi{}^T+\lambda\mathbf{I}_N)^{-1}\mathbf{Y} \tag{9}$$

By defining a Mercer kernel matrix $\mathbf{K}$, we have

$$\mathbf{K}=\mathbf{X}_\phi\mathbf{X}_\phi{}^T\in R^{N\times N} \tag{10}$$

Therefore, for an unseen sample $\mathbf{x}$, its prediction can be expressed as

$$\begin{aligned}f(\mathbf{x})=&\phi(\mathbf{x})\mathbf{A}=\phi(\mathbf{x})\mathbf{X}_\phi{}^T(\mathbf{X}_\phi\mathbf{X}_\phi{}^T+\lambda\mathbf{I}_N)^{-1}\mathbf{Y}\\=&\begin{bmatrix}K(\mathbf{x},\mathbf{x}_1)\\...\\K(\mathbf{x},\mathbf{x}_N)\end{bmatrix}^T(\mathbf{K}+\lambda\mathbf{I}_N)^{-1}\mathbf{Y}\end{aligned} \tag{11}$$

where K denotes the kernel function. By substituting (5) to (1), we find that the loss function in the reproducing kernel Hilbert space can be formulated as $\left\|\mathbf{K}\boldsymbol{\alpha}-\mathbf{Y}\right\|_F^2$, where $\mathbf{K}$ is a Mercer kernel matrix. Therefore, because of the diversity of kernels, we can redesign the representation of $\mathbf{K}$ to achieve multi-kernel learning.

III. NS-RLS-MKL

In this section, we first define the objective function of our proposed model NS-RLS-MKL. Then we deduce the solution to the objective function and list the detailed algorithm steps.

A. Non-Sparse Multi-Kernel Regression

Multi-Kernel learning attempts to obtain better mapping performance by combining different kernel functions or kernel functions with different kernel parameters. There are many ways to combine kernel functions, among which the linear combination is the most used. Suppose that we have *M* different kernel functions, $\mathbf{K}_z$ represents the *z-th* one, $1\le z\le Z$, then a linear combination of these kernel functions can be expressed as

$$K=\sum_{z=1}^{Z}\theta_z\mathbf{K}_z \tag{12}$$

where θ_z is the combination coefficient. How to learn θ_z can determine the utilization of pattern information

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

contained in different kernels. For example, if l_1-norm is imposed on the learning of θ_z, we will have a sparse distribution of θ_z, which means only most of discriminant kernels will be used. If l_p-norm (p>1) is imposed, we will have s non-sparse distribution of θ_z, which means pattern information contained in all kernels will be used. Typically, we have a basic assumption that all kernels contain complementary patterns. It is important to use and fuse such complementary information effectively. Therefore, in this study, we use l_p-norm (p>1) to impose on the learning of θ_z to generate non-sparse distribution so that the complementary information contained in each kernel can be fully exploited. Thus, the multi-kernel regression with the l_p-norm regularization can be formulated as

$$\begin{aligned} &\min_{\mathbf{A}} \left\| \sum_{z=1}^{Z} \theta_z \mathbf{K}_z \mathbf{A} - \mathbf{Y} \right\|_F^2 + \lambda \|\mathbf{A}\|_F^2 \\ &\text{s.t} \sum_{z=1}^{Z} (\theta_z)^p \le 1, p>1 \end{aligned} \tag{13}$$

B. Regularized Label Softening

When the non-sparse multi-kernel regression model shown in (12) is applied for multi-class classification tasks, $\mathbf{Y}$ must be a strict binary label matrix. Previous studies showed that excessively fitting a strict binary label matrix cannot learn a discriminant model. To solve this problem, similar to [37], we introduce two matrices $\mathbf{D}$ and $\mathbf{M}$ to soften the label matrix $\mathbf{Y}$ so as to enlarge the margins between classes. The softening process is formulated as

$$\tilde{\mathbf{Y}}=\mathbf{Y}+\mathbf{D}\odot\mathbf{M} \tag{14}$$

where $\odot$ denotes the Hadamard operator, $\mathbf{D}$ and $\mathbf{M}$ are defined as follows,

$$d_{ij} = \begin{cases} +1 & \text{if } y_{ij}=1 \\ -1 & \text{if } y_{ij}=0 \end{cases} \tag{15}$$

$$\mathbf{M}=\begin{bmatrix} m_{11} & \cdots & m_{1C} \\ \vdots & m_{ij} & \vdots \\ m_{N1} & \cdots & m_{NC} \end{bmatrix} \quad i=1,2,...,N,\ j=1,2,...,C,\ m_{ij}\ge 0 \tag{16}$$

After label softening, the original multi-kernel regression model in (12) can be updated as

$$\begin{aligned} &\min_{\mathbf{A}} \left\| \sum_{z=1}^{Z} \theta_z \mathbf{K}_z \mathbf{A} - (\mathbf{Y}+\mathbf{D}\odot\mathbf{M}) \right\|_F^2 + \lambda \|\mathbf{A}\|_F^2 \\ &\text{s.t} \sum_{z=1}^{Z} (\theta_z)^p \le 1, p>1 \end{aligned} \tag{17}$$

Although label softening can enlarge the class margins and hence improve classification performance, the overfitting will occur owning to the freedom of fitting. Therefore, overfitting should also be suppressed in the pursuit of more discriminative models. To this end, based on manifold learning, we use a regularized term to control data fitting. This regularized term is designed permitted on an assumption that samples in the same class should be kept as close as possible when they are transformed into the label space. To specify this regularized term, we first construct an undirected graph to capture the relationships between samples, which is defined as

$$g_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma}} & \mathbf{x}_i,\ \mathbf{x}_j \text{ are in the same class} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where σ is the kernel width. From the above equation, we see that if two samples are in the same class, the closer the distance, the bigger the weight. If they are in different classes, the weight is 0. Therefore, when samples are transformed into label space, we can use the following objective to insure our assumption,

$$\min_{f} \sum_{ij} \left\| f(\mathbf{x}_i) - f(\mathbf{x}_j) \right\|_F^2 g_{ij} \tag{19}$$

where $f(\mathbf{x}_i)$ is the decision function in the multi-kernel feature space that can be expressed as $\sum_{z=1}^{Z} \theta_z \mathbf{K}_z(\mathbf{x}, \mathbf{x}_i)\mathbf{A}$. Thus, the manifold regularized term can be changed into

$$\begin{aligned} &\min_{\mathbf{A}} \sum_{ij} \left\| f(\mathbf{x}_i) - f(\mathbf{x}_j) \right\|_F^2 g_{ij} = \\ &\min_{\mathbf{A}} \operatorname{tr}(\mathbf{A}^T \sum_{z=1}^{Z} \theta_z \mathbf{K}_z \mathbf{L} \sum_{z=1}^{Z} \theta_z \mathbf{K}_z \mathbf{A}) \end{aligned} \tag{20}$$

where $\mathbf{L}$ is the Laplace matrix that can be computed by $\mathbf{L}=\mathbf{T}-\mathbf{G}$, $\mathbf{Z}$ is a diagonal matrix in which each element can be computed by $t_{ii} = \sum_j g_{ij}$. By substituting (20) into (17), we have our final non-sparse multi-kernel regression model,

$$\begin{aligned} &\min_{\mathbf{A}} \left\| \left(\sum_{z=1}^{Z} \theta_z \mathbf{K}_z \right) \mathbf{A} - (\mathbf{Y}+\mathbf{D}\odot\mathbf{M}) \right\|_F^2 + \lambda \|\mathbf{A}\|_2^2 \\ &\qquad + \gamma \operatorname{tr}\left(\mathbf{A}^T \left(\sum_{z=1}^{Z} \theta_z \mathbf{K}_z \right) \mathbf{L} \left(\sum_{z=1}^{Z} \theta_z \mathbf{K}_z \right) \mathbf{A} \right) \\ &\text{s.t} \sum_{z=1}^{Z} (\theta_z)^p \le 1, p>1 \end{aligned} \tag{21}$$

C. Optimization

Regarding our final objective, there are three model parameters, i.e., α_m, $\mathbf{A}$ and $\mathbf{M}$ should be optimized. The following three theorems are proposed for handling such optimization problem.

Theorem 1: When θ_z and $\mathbf{M}$ are fixed, the objective function converges to its minimum if only if $\mathbf{A}=(\tilde{\mathbf{K}}^T\tilde{\mathbf{K}}+\gamma\tilde{\mathbf{K}}^T\mathbf{L}\tilde{\mathbf{K}}+\lambda\mathbf{I}_N)^{-1}\tilde{\mathbf{K}}^T(\mathbf{Y}+\mathbf{D}\odot\mathbf{M})$, where $\sum_{z=1}^{Z} \theta_z \mathbf{K}_z = \tilde{\mathbf{K}}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Proof: When θ_z and $\mathbf{M}$ are fixed, suppose that $\sum_{z=1}^{Z}\alpha_z\mathbf{K}_z=\tilde{\mathbf{K}}$, then the optimization problem becomes

$$\begin{aligned} J(\mathbf{A}) = \min_{\mathbf{A}} &\left\|\tilde{\mathbf{K}}\mathbf{A}-(\mathbf{Y}+\mathbf{D}\odot\mathbf{M})\right\|_F^2 \\ &+\lambda\|\mathbf{A}\|_F^2+\gamma\mathrm{tr}\left(\mathbf{A}^T\tilde{\mathbf{K}}\mathbf{L}\tilde{\mathbf{K}}\mathbf{A}\right) \end{aligned} \tag{22}$$

By setting the derivative partial of $J(\mathbf{A})$ w.r.t $\mathbf{A}$ to 0, that is, $\partial J(\mathbf{A})/\partial\mathbf{A}=0$

$$\begin{aligned} &\Rightarrow 2\tilde{\mathbf{K}}^T\tilde{\mathbf{K}}-2\tilde{\mathbf{K}}^T((\mathbf{Y}+\mathbf{D}\odot\mathbf{M})) \\ &\qquad +2\lambda\mathbf{A}+2\gamma\tilde{\mathbf{K}}^T\mathbf{L}\tilde{\mathbf{K}}\mathbf{A}=0 \\ &\Rightarrow \mathbf{A}=(\tilde{\mathbf{K}}^T\tilde{\mathbf{K}}+\gamma\tilde{\mathbf{K}}^T\mathbf{L}\tilde{\mathbf{K}}+\lambda\mathbf{I}_N)^{-1}\tilde{\mathbf{K}}^T((\mathbf{Y}+\mathbf{D}\odot\mathbf{M})) \end{aligned} \tag{23}$$

So, the proof of **Theorem 1** is achieved**.**

Theorem 2: When θ_z and $\mathbf{A}$ are fixed, the objective function converges to its minimum if only if $\mathbf{M}=\max(\mathbf{D}\odot\mathbf{E},0)$, where $\mathbf{E}=\left(\sum_{z=1}^{Z}\theta_z\mathbf{K}_z\right)\mathbf{A}-\mathbf{Y}$.

Proof: When θ_z and $\mathbf{A}$ are fixed, suppose that $\left(\sum_{z=1}^{Z}\theta_z\mathbf{K}_z\right)\mathbf{A}-\mathbf{Y}=\mathbf{E}$, then the optimization problem becomes

$$\begin{aligned} &J(\mathbf{M}) = \min_{\mathbf{M}}\left\|\mathbf{E}-\mathbf{D}\odot\mathbf{M}\right\|_2^2 \\ &\text{s.t } m_{ij}\geq 0 \end{aligned} \tag{24}$$

We have the fact that the squared l_2-norm of matrix can be decoupled element by element. Therefore, the optimization problem in (23) can be decomposed into $N\times C$ subproblems. Regarding element m_{ij} of $\mathbf{M}$, the corresponding subproblem can be expressed as

$$\begin{aligned} &J(m_{ij}) = \min_{\mathbf{M}}\left\|e_{ij}-d_{ij}m_{ij}\right\|_2^2 \\ &\text{s.t } m_{ij}\geq 0 \end{aligned} \tag{25}$$

where e_{ij} and d_{ij} denote the i-th row and j-th column element in matrices $\mathbf{E}$ and $\mathbf{M}$. Owning to $(d_{ij})^2=1$, so $(e_{ij}-d_{ij}m_{ij})^2=(d_{ij}m_{ij}-e_{ij})^2$ holds. Additionally, due to $m_{ij}\geq 0$, we have $m_{ij}=\max(d_{ij}e_{ij},0)$. Therefore, regarding $\mathbf{M}$, we have its finally solution as follows,

$$\mathbf{M}=\max(\mathbf{D}\odot\mathbf{E},0) \tag{26}$$

Theorem 3: When $\mathbf{M}$ and $\mathbf{A}$ are fixed, the objective function converges to its minimum if only if $\theta_z=(\mathbf{A}^T\mathbf{K}_z\mathbf{A})^{\frac{2}{p+1}}(\sum_{z'=1}^{Z}(\mathbf{A}^T\mathbf{K}_{z'}\mathbf{A})^{\frac{2p}{p+1}})^{\frac{-1}{p}}$.

Proof: According to [38], when $\mathbf{M}$ and $\mathbf{A}$ are fixed, the final optimization problem of θ_z becomes

$$J(\theta_z) = \min_{\theta_z}\frac{1}{2}\mathbf{A}^T\sum_{z=1}^{Z}\frac{\mathbf{K}_z}{\theta_z}\mathbf{A}+\frac{\eta}{2}\sum_{z=1}^{Z}(\theta_z)^p \tag{27}$$

By setting the partial derivative of $J(\theta_z)$ w.r.t θ_z to 0, we have

$$-\frac{1}{2}\mathbf{A}^T\sum_{z=1}^{Z}\frac{\mathbf{K}_z}{(\theta_z)^2}\mathbf{A}+\eta(\theta_z)^{p-1}\|\boldsymbol{\theta}\|_p^{2-p}=0 \tag{28}$$

Therefore, we have the following optimality condition,

$$\exists\xi\ \forall z=1,2,...,M:\ \theta_z=\xi(\mathbf{A}^T\mathbf{K}_z\mathbf{A})^{\frac{2}{p+1}} \tag{29}$$

Because $\mathbf{A}^T\mathbf{K}_z\mathbf{A}\neq 0$, according the same argument as in the proof of Theorem 1 in [38], the constraint $\sum_{z=1}^{Z}(\theta_z)^p\leq 1$ in (21) is at the upper bound, that is to say, $\sum_{z=1}^{Z}(\theta_z)^p=1$ holds for an optimal θ. So, $\xi=(\sum_{z'=1}^{Z}(\mathbf{A}^T\mathbf{K}_{z'}\mathbf{A})^{\frac{2p}{p+1}})^{\frac{-1}{p}}$. Therefore, the proof of **Theorem 3** is achieved.

D. Algorithm

The detailed algorithm steps of the proposed model are listed in Algorithm1.

Algorithm 1: **NS-RLS-MKL**

Input: Training data $\{\mathbf{x}_i,y_i\}_{i=1}^N$, regularized parameters λ and γ

Output: θ_z, $\mathbf{A}$, and $\mathbf{M}$

Procedure:

[1]Transform the label vector $[y_1,y_2,...,y_N]^T$ to a binary label matrix.

[2]Construct the Laplace matrix $\mathbf{L}$ according to the training data.

[3]Initialize $\mathbf{M}^{(0)}$ and $\theta_z^{(0)}$.

[4]$t\leftarrow 0$.

Repeat

[5]$t\leftarrow t+1$

[6]Using $\mathbf{A}=(\tilde{\mathbf{K}}^T\tilde{\mathbf{K}}+\gamma\tilde{\mathbf{K}}^T\mathbf{L}\tilde{\mathbf{K}}+\lambda\mathbf{I}_N)^{-1}\tilde{\mathbf{K}}^T(\mathbf{Y}+\mathbf{D}\odot\mathbf{M})$ to update $\mathbf{A}^{(t)}$.

[7]Using $\mathbf{M}=\max(\mathbf{D}\odot\mathbf{E},0)$ to update $\mathbf{M}^{(t)}$.

[8]Using $\theta_z=(\mathbf{A}^T\mathbf{K}_z\mathbf{A})^{\frac{2}{p+1}}(\sum_{z'=1}^{Z}(\mathbf{A}^T\mathbf{K}_{z'}\mathbf{A})^{\frac{2p}{p+1}})^{\frac{-1}{p}}$ to update $\theta_z^{(t)}$

Until $(\left\|\mathbf{A}^{(t)}-\mathbf{A}^{(t-1)}\right\|_F^2<\zeta)$

From Algorithm 1, we find that the time complexity of NS-RLS-MKL mainly consists of updating $\mathbf{A}$, updating $\mathbf{M}$ and updating θ_z. Due to the matrix inversion, the asymptotic time complexity of updating $\mathbf{A}$ is $O(N^3)$. The asymptotic time complexity of updating $\mathbf{M}$ is $O(N)$. The asymptotic time complexity of updating θ_z is $O(N^2)$. Hence, the final asymptotic time complexity of NS-RLS-MKL can be approximated to $O(N^3)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

IV. Experimental Studies

In this section, a synthetic multimodal dataset and several UCI multimodal datasets are introduced for performance evaluation. Additionally, to highlight the proposed model, several benchmarking models are introduced for comparison.

A. Datasets

We generate a 3-modal dataset containing 600 samples which is shown in Fig.1. Each modality is derived from Fig.1(a) by projecting onto different coordinate planes. This synthetic dataset simulates different modalities of data collected by different sensors on the same object.

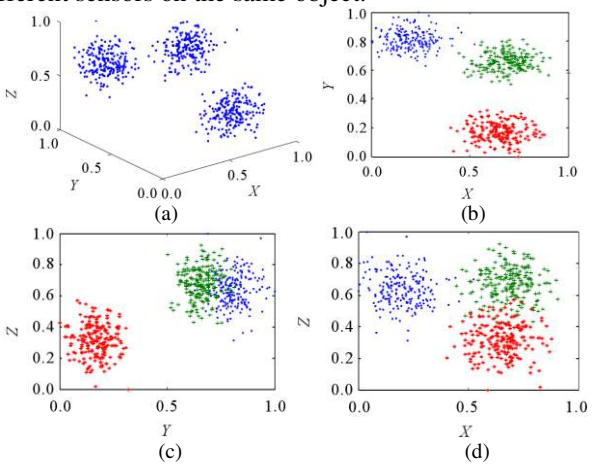

Fig.1 Three-modal synthetic dataset. (a) Orginal data. (b) The first modality (Mod-XY). (c) The second modality (Mod-YZ). (d) The three modality (Mod-XZ).

Additionally, 3 UCI multimodal datasets, MF (Multiple Features), IS (Image Segmentation) and WTP (Water Treatment Plant) are introduced for performance evaluation. Table 1 lists the detailed information of these datasets.

Table I. Introduction of multimodal datasets

Datasets	Modalities	Description of each modality	#Features	Size
MF	Mfeat-fou	Mfeat-fou contains 76 Fourier coefficients of the character shapes	76	2000
	Mfeat-fac	Mfeat-fac contains 216 profile correlations	216	
	Mfeat-kar	Mfeat-kar contains 64 Karhunen-Love coefficients	64	
	Mfeat-pix	Mfeat-pix contains 240 pixel averages in 2 x 3 windows	240	
	Mfeat-zer	Mfeat-zer contains 47 Zernike moments	47	
	Mfeat-mor	Mfeat-mor contains 6 morphological variables	6	
IS	Shape	Shape contains 9 features about the shape information of the 7 images	9	2310
	RGB	RGB contains 10 features about the RGB values of the 7 images	10	
WTP	Input	Input contains the first 22 features describing different input conditions	22	527
	Output	Output contains the 23th-29th features describing output demands	7	
	Performance input	Performance input contains the 30th-34th features describing performance input demands	5	
	Global performance input	Global performance input contains the 35th-38th features describing global performance input demands	4	

B. Settings

To highlight the classification performance of the proposed model, we introduce Ridge as the baseline and MV-L2-SVM and MV-TSK-FS as benchmarking models.

Ridge: It is taken as the baseline. We perform it on each modality and record the corresponding classification accuracy. The overall accuracy on all modalities is computed by linearly combining the accuracy on each modality. The combination coefficient can be used as the reciprocal of the training error. The regularized parameter in Ridge is determined by 10 cross-validation (10-CV) from [0.01, 0.02, …, 10].

MV-L2-SVM: It uses L2-SVM as the basic component and view (modality)-consistence as the multimodal learning strategy. There are two parameters should be set in advance. The kernel width is determined by 10-CV from [s/256, s/128, …, 256s], where s is the mean norm of the training data. The penalty parameter is determined by 10-CV from [10^0, 10^1, …, 10^7].

MV-TSK-FS: It uses the 1-order TSK fuzzy system as the basic component and utilizes view (modality)-consistence and view-weighting as two multimodal learning strategies. There are four parameters should be set in advance. The number of fuzzy rules is determined by 10-CV from [5, 10, …, 30]. The three regularized parameters are determined by 10-CV from [10^{-3}, 10^{-2}, …, 10^3].

Regarding the proposed model, we adopt the same kernel combination strategy "all-single" used in [1] to combine modalities and kernels, as shown in Fig.2. There are three parameters should be set in advance. The regularized parameter λ is determined by 10-CV from [0.01, 0.02, …, 10]. The regularized parameter γ is determined by 10-CV from [10^{-3}, 10^{-2}, …, 10^3].

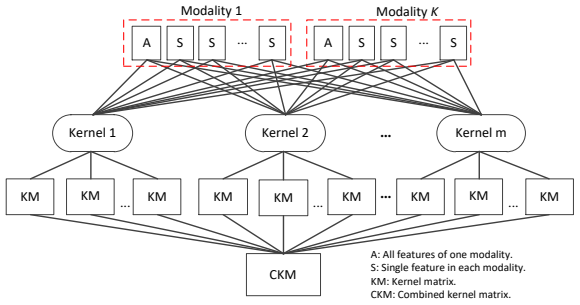

Fig.2 "All-single" kernel and modality combination strategy.

For all models, we use *Accuracy* and *F1-measure* to quantificationally evaluate their performance. *Accuracy* is defined as the ratio of the number of corrected classified samples to the total number of samples. *F1-measure* is defined as $2\times P\times R/(P+R)$, where $P=TP/(TP+FP)$, $R=TP/(TP+FN)$. TP represents true positives, FP represents false positives, and FN represents false negatives.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

C. Experimental results

Table II and Table III show the classification performance in terms of *Accuracy* and *F1-measure*, respectively. "+" means that the performance improvement of NS-RLS-MKL is significant at 5% level when comparing with the corresponding model. The best performance is marked in bold. Results in the parentheses are the standard deviation. It observes that our proposed model NS-RLS-MKL wins the best performance both in single modality and multiple modalities. In particular, in the multimodal case, the *Accuracy* reaches 1, which is better than that of the view weighting strategy used in MV-L2-SVM and MV-TSK-FS.

Table II. Classification performance in terms of *Accuracy* and corresponding standard deviation on synthetic dataset

Modalities	Ridge	MV-L2-SVM	MV-TSK-FS	NS-RLS-MKL
Mod-XY	0.9756 (0.0168)	0.9672$^{(+)}$ (0.0221)	0.9761 (0.0162)	**0.9983** (0.0027)
Mod-YZ	0.6906$^{(+)}$ (0.0251)	0.7089$^{(+)}$ (0.0314)	0.6828$^{(+)}$ (0.0384)	**0.8383** (0.0284)
Mod-XZ	0.9061$^{(+)}$ (0.0169)	0.8967$^{(+)}$ (0.0270)	0.6794$^{(+)}$ (0.0250)	**0.9744** (0.0146)
Full	0.9739 (0.0114)	0.9406$^{(+)}$ (0.0399)	0.9644$^{(+)}$ (0.0200)	**1.0000** (0)

Table III. Classification performance in terms of *F1-measure* and corresponding standard deviation on synthetic dataset

Modalities	Ridge	MV-L2-SVM	MV-TSK-FS	NS-RLS-MKL
Mod-XY	0.9640$^{(+)}$ (0.0252)	0.9541$^{(+)}$ (0.0311)	0.9671$^{(+)}$ (0.0204)	**0.9977** (0.0037)
Mod-YZ	0.6314$^{(+)}$ (0.0288)	0.6623$^{(+)}$ (0.0291)	0.6327$^{(+)}$ (0.0457)	**0.7727** (0.0395)
Mod-XZ	0.8730$^{(+)}$ (0.0235)	0.8633$^{(+)}$ (0.0275)	0.6238$^{(+)}$ (0.0277)	**0.9498** (0.0160)
Full	0.9623$^{(+)}$ (0.0170)	0.9161$^{(+)}$ (0.0531)	0.9474$^{(+)}$ (0.0292)	**1.0000 (0)**

The advantages of classification performance can also be observed from the results on real-life datasets, as shown in Table IV and Table V. To be specific, on dataset MF, except the modality Mfeat-fac, our proposed model NS-RLS-MKL wins the best on all single modality and multiple modalities. On data set IS, our proposed model NS-RLS-MKL wins the best on all single modality and multiple modalities. Especially, the classification performance is improved by 10% compared with the baseline. On dataset WTP, except the modality Output and modality Global performance input, our proposed model NS-RLS-MKL wins the best on all single modality and multiple modalities. We also have the similar conclusion when *F1-measure* is adopted as the criterion.

Unlike MV-L2-SVM and MV-TSK-FS which both adopt weighting learning as their fusion strategy, the proposed model combines all modality features and use different kernels to mine the complementary pattern information. Moreover, non-sparse coefficient can help maintain more enough patterns involved in different modalities. Additionally, in our multi-classification task, we use label softening strategy to enlarge the margins between classes to guarantee promising performance and manifold-based regularization to anti overfitting problem. That is why our model performs much better than Ridge, MV-L2-SVM and MV-TSK-FS.

Table IV. Classification performance in terms of *Accuracy* and corresponding standard deviation on UCI datasets

Datasets	**Modalities**	**Ridge**	**MV-L2-SVM**	**MV-TSK-FS**	**NS-RLS-MKL**
MF	Mfeat-fac	0.7958 (0.0138)	**0.8002** (0.0111)	0.7986 (0.0057)	0.7855 (0.0176)
	Mfeat-fou	0.7590 (0.0095)	0.7585 (0.0142)	0.7552 (0.0162)	**0.7655** (0.0128)
	Mfeat-kar	0.7798$^{(+)}$ (0.0236)	0.7728$^{(+)}$ (0.0156)	0.7707$^{(+)}$ (0.0160)	**0.8368** (0.0200)
	Mfeat-mor	0.5812$^{(+)}$ (0.0253)	0.5507$^{(+)}$ (0.0188)	0.5713$^{(+)}$ (0.0073)	**0.6748** (0.0191)
	Mfeat-pix	0.8107$^{(+)}$ (0.0163)	0.8028$^{(+)}$ (0.0217)	0.8110$^{(+)}$ (0.0189)	**0.8425** (0.0164)
	Mfeat-zer	0.2488 (0.0173)	0.2325 (0.0131)	0.2412 (0.0141)	**0.2673** (0.0132)
	Full	0.9210$^{(+)}$ (0.0081)	0.9252$^{(+)}$ (0.0060)	0.9255$^{(+)}$ (0.0143)	**0.9512** (0.0098)
IS	Shape	0.3915$^{(+)}$ (0.0235)	0.3879$^{(+)}$ (0.0090)	0.3973$^{(+)}$ (0.0125)	**0.6437** (0.0166)
	RGB	0.6730$^{(+)}$ (0.0142)	0.6771$^{(+)}$ (0.0161)	0.6740$^{(+)}$ (0.0120)	**0.8162** (0.0168)
	Full	0.8232$^{(+)}$ (0.0235)	0.8215$^{(+)}$ (0.0170)	0.8176$^{(+)}$ (0.0151)	**0.9341** (0.0133)
WTP	Input	0.5215 (0.0505)	0.5025 (0.0604)	0.5329 (0.0276)	**0.5601** (0.0442)
	Output	**0.4468**$^{(+)}$ (0.0408)	0.4196 (0.0401)	0.4449 (0.0242)	0.4025 (0.0260)
	Performance input	0.4671 (0.0378)	0.4620 (0.0500)	0.4665 (0.0347)	**0.4753** (0.0443)
	Global performance input	**0.3867** (0.0216)	0.3753 (0.0358)	0.3861 (0.0209)	0.3747 (0.0268)
	Full	0.5184 (0.0429)	0.5253 (0.0504)	0.5348 (0.0342)	**0.5494** (0.0452)

Table V. Classification performance in terms of *F1-measure* and corresponding standard deviation on UCI datasets

	Modalities	Ridge	MV-L2-SVM	MV-TSK-FS	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3087738, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

MF	Mfeat-fac	0.4363 (0.0355)	0.4423 (0.0255)	0.4158$^{(+)}$ (0.0221)	**0.4489** (0.0453)
	Mfeat-fou	0.4605 (0.0401)	0.4444 (0.0217)	0.4294$^{(+)}$ (0.0277)	**0.4682** (0.0427)
	Mfeat-kar	0.4471$^{(+)}$ (0.0356)	0.4434$^{(+)}$ (0.0194)	0.4342$^{(+)}$ (0.0265)	**0.5272** (0.0370)
	Mfeat-mor	0.3213$^{(+)}$ (0.0313)	0.3002$^{(+)}$ (0.0208)	0.3144$^{(+)}$ (0.0217)	**0.3783** (0.0265)
	Mfeat-pix	0.4550$^{(+)}$ (0.0384)	0.4438$^{(+)}$ (0.0490)	0.4568$^{(+)}$ (0.0481)	**0.5005** (0.0514)
	Mfeat-zer	0.0258$^{(+)}$ (0.0010)	0.0194$^{(+)}$ (0.0002)	0.0223$^{(+)}$ (0.0007)	**0.0841** (0.0087)
	Full	0.7216$^{(+)}$ (0.0257)	0.7182$^{(+)}$ (0.0261)	0.7331$^{(+)}$ (0.0425)	**0.8089** (0.0231)
IS	Shape	0.1025$^{(+)}$ (0.0123)	0.1315$^{(+)}$ (0.0087)	0.1736$^{(+)}$ (0.0056)	**0.2536** (0.0188)
	RGB	0.4667$^{(+)}$ (0.0236)	0.4608$^{(+)}$ (0.0211)	0.4655$^{(+)}$ (0.0218)	**0.6097** (0.0314)
	Full	0.6127$^{(+)}$ (0.0446)	0.6052$^{(+)}$ (0.0396)	0.6022$^{(+)}$ (0.0202)	**0.8101** (0.0392)
WTP	Input	0.5386$^{(+)}$ (0.0546)	0.5383$^{(+)}$ (0.0461)	0.5406$^{(+)}$ (0.0366)	**0.5858** (0.0391)
	Output	0.5429 (0.0316)	0.5015 (0.0458)	0.5348 (0.0237)	**0.5500** (0.0199)
	Performance input	0.5284 (0.0262)	0.5258 (0.0396)	**0.5396** (0.0260)	0.5385 (0.0364)
	Global performance input	**0.5417** (0.0272)	0.5229 (0.0392)	0.5355 (0.0235)	0.5396 (0.0282)
	Full	0.5357$^{(+)}$ (0.0384)	0.5251$^{(+)}$ (0.0403)	0.5303$^{(+)}$ (0.0321)	**0.5838** (0.0377)

D. A case study

To evaluate the application ability of the proposed model NS-RLS-MKL in remoting sensor data fusion, we select two types of remoting sensor data. One is high resolution dataset NWPUVHR (Northwestern Poly-Technique University and Very-high-resolution Remote Sensing Images) [39] and another is low resolution dataset UCMLU (University of California Merced Land Use) [40], as shown in Fig.3.

Modality-1: high resolution

Tennis court Warehouse Golf field Baseball field

Modality-2: low resolution

Tennis court Warehouse Golf field Baseball field

Fig.3 Samples of multimodal dataset

We select 1000 samples including four classes from each modality in our case study. We use our previous deep feature extractor [41] to extract 1024 deep features from each remoting sensor image. Then we use the embedded-based feature selection method [42] to select 15 features from the 1024 deep features from each modality for NS-RLS-MKL. Table VI and Table VII list the final classification performance in terms of *Accuracy* and *F1-measure*.

From the results, it observes that NS-RLS-MKL always performs the best among the comparison models and baseline. From Fig.3, we see that the high-resolution remoting sensor images provide very detailed information (local information) while the low-resolution remoting sensor images provide rough information (global information). In our proposed model, we combine them and use different kernels associating with non-spare kernel coefficient to mine the complementary pattern information from local information and global information. However, unlike our proposed model, the comparison ones only determine which modality is more important and assign a bigger weight to the important one. Such combination strategy does not fully utilize the deep correlation across different modalities.

Table VI. Classification performance in terms of *Accuracy* and corresponding standard deviation on remoting sensor data

Modalities	Ridge	MV-L2-SVM	MV-TSK-FS	NS-RLS-MKL
High resolution	0.9024 (0.0036)	0.8965$^{(+)}$ (0.024)	0.9024 (0.0063)	**0.9123** (0.0018)
Low resolution	0.7784 (0.0078)	0.7432$^{(+)}$ (0.0024)	0.7641 (0.0021)	**0.7896** (0.0101)
Full	0.9139$^{(+)}$ (0.0014)	0.9009$^{(+)}$ (0.0099)	0.9130$^{(+)}$ (0.0110)	**0.9325** (0.0074)

Table VII. Classification performance in terms of *F1-measure* and corresponding standard deviation on remoting sensor data

Modalities	Ridge	MV-L2-SVM	MV-TSK-FS	NS-RLS-MKL
High resolution	0.8748$^{(+)}$ (0.0280)	0.8852$^{(+)}$ (0.0181)	0.8957 (0.0029)	**0.9028** (0.0069)
Low resolution	0.7401 (0.0098)	0.7369$^{(+)}$ (0.0078)	0.7489 (0.0038)	**0.7698** (0.0058)
Full	0.9157 (0.0140)	0.9087$^{(+)}$ (0.0057)	0.9068$^{(+)}$ (0.0036)	**0.9258** (0.0074)

V. CONCLUSIONS

How to effectively fuse different modal data and mine the hidden value of data through the complementary information between modalities is the main concern of big data research at

the present stage. In this study, we proposed a novel multi-kernel model for multimodal data fusion based non-sparse multi-kernel learning. In classification scenarios, we introduce a positive matrix to soften the binary label matrix so that the margins between classes are enlarged as much as possible. Therefore, the label fitting becomes freer. Even the free label fitting may cause overfitting, by using the manifold-based regularization, this problem is solved to the maximum extent possible. Additionally, our proposed model is derived from non-sparse multi-kernel learning, which is better suited for multimodal data fusion than sparse based. We generate a synthetic multimodal dataset and introduce several real-life multimodal datasets to demonstrate the advantages of the proposed model. The comparison results with baseline and other multimodal models show that our model performs better.

However, our model also faces a challenge that when the dimension of the input feature space is very high, it will consume a lot of memory spaces to store the kernels. In our future work, we will develop more effective optimization method to solve this problem.

REFERENCE

[1] Y. Zhang, S. Wang, K. Xia, et al. “Alzheimer’s disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion,” Information Fusion, vol, 66, pp. 170-183, 2021.

[2] D. Hong, N. Yokoya, J. Chanussot, X. Zhu, "An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1923-1938, 2019

[3] D. Hu et al., "Disentangled-Multimodal Adversarial Autoencoder: Application to Infant Age Prediction With Incomplete Multimodal Neuroimages," in IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 4137-4149, Dec. 2020.

[4] R. Wang, X. -J. Wu and J. Kittler, "Graph Embedding Multi-Kernel Metric Learning for Image Set Classification With Grassmannian Manifold-Valued Features," in IEEE Transactions on Multimedia, vol. 23, pp. 228-242, 2021.

[5] Y. Zheng, X. Wang, G. Zhang, B. Xiao, F. Xiao and J. Zhang, "Multi-Kernel Coupled Projections for Domain Adaptive Dictionary Learning," in IEEE Transactions on Multimedia, vol. 21, no. 9, pp. 2292-2304, Sept. 2019.

[6] S. M. Xiang, F. P. Nie, G. F. Meng, C. H. Pan, and C. S. Zhang, “Discriminative least squares regressions for multiclass classification and feature selection,” IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[7] F. P. Nie, H. Wang, H. Huang, and C. Ding, “Adaptive loss minimization for semi-supervised elastic embedding,” in Proc. 23rd Int. Joint Conf. Artif. Intell., Beijing, China, 2013, pp. 1565–1571.

[8] X. Fang, Y. Xu, X. Li, Z. Lai, W. K. Wong and B. Fang, "Regularized Label Relaxation Linear Regression," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 4, pp. 1006-1018, April 2018.

[9] X. Zhang, Q. Song, R. Liu, W. Wang and L. Jiao, "Modified Co-Training With Spectral and Spatial Views for Semisupervised Hyperspectral Image Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp. 2044-2055, June 2014.

[10] C. M. Nguyen, X. Li, R. D. Blanton and X. Li, "Partial Bayesian Co-training for Virtual Metrology," in IEEE Transactions on Industrial Informatics, vol. 16, no. 5, pp. 2937-2945, May 2020.

[11] Y. Xu, L. Qin and Q. Huang, "Coupling Reranking and Structured Output SVM Co-Train for Multitarget Tracking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 6, pp. 1084-1098, June 2016.

[12] B. -z. Zhang and W. -l. Zuo, "Co-EM Support Vector Machine Based Text Classification from Positive and Unlabeled Examples," 2008 First International Conference on Intelligent Networks and Intelligent Systems, Wuhan, China, 2008, pp. 745-748.

[13] Z. Zhou and M. Li, "Semisupervised Regression with Cotraining-Style Algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 11, pp. 1479-1493, Nov. 2007.

[14] Sun S, Jin F. Robust co-training”. International Journal of Pattern Recognition & Artificial Intelligence, vol. 25, no. 7, pp. 1113-1126, 2011.

[15] Yu S, Krishnapuram B, Rosales R, et al. “Bayesian co-training,” Journal of Machine Learning Research, vol. 12, no. 3, pp. 2649-2680, 2011.

[16] Zhao X, Evans N, Dugelay J L. “A subspace co-training framework for multi-view clustering,” Pattern Recognition Letters, vol. 14, no. 1, pp. 73-82, 2014.

[17] B. Fan, Y. Cong, J. Tian and Y. Tang, "Reliable Multi-Kernel Subtask Graph Correlation Tracker," in IEEE Transactions on Image Processing, vol. 29, pp. 8120-8133, 2020.

[18] J. Huang et al., "Identifying Resting-State Multifrequency Biomarkers via Tree-Guided Group Sparse Learning for Schizophrenia Classification," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 1, pp. 342-350, Jan. 2019.

[19] W. Gao, J. Huang and J. Han, "Multi-channel differencing adaptive noise cancellation with multi-kernel method," in Journal of Systems Engineering and Electronics, vol. 26, no. 3, pp. 421-430, June 2015.

[20] R. Xu, J. Hu, Q. Lu, D. Wu and L. Gui, "An ensemble approach for emotion cause detection with event extraction and multi-kernel SVMs," in Tsinghua Science and Technology, vol. 22, no. 6, pp. 646-659, December 2017.

[21] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016, pp. 439-448.

[22] H. Zheng et al. “A novel multiple kernel sparse representation based classification for face recognition,” KSII Transactions on Internet & Information Systems, vol.8, no.4, pp. 1463-2664, 2014.

[23] D. R. Hardoon et al. “An overview with application to learning methods,” Neural Computation, vol. 16, no. 12, pp. 2639-2664, 2014.

[24] J. Costa Pereira et al., "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 521-535, March 2014.

[25] M. Kan, S. Shan, H. Zhang, S. Lao and X. Chen, "Multi-View Discriminant Analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 188-194, 1 Jan. 2016.

[26] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang and J. Jiang, "Sparse Unsupervised Dimensionality Reduction for Multiple View Data," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 10, pp. 1485-1496, Oct. 2012.

[27] H. Liu, M. Shao and Y. Fu, "Feature Selection with Unsupervised Consensus Guidance," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2319-2331, 1 Dec. 2019.

[28] Y. Jiang et al. “Semi-supervised unified latent factor learning with multi-view data,” vol. 25, no.7, pp.1635-1645, 2014.

[29] G. Ding, Y. Guo and J. Zhou, "Collective Matrix Factorization Hashing for Multimodal Data," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 2083-2090.

[30] M. M. Kalayeh, H. Idrees and M. Shah, "NMF-KNN: Image Annotation Using Weighted Multi-view Non-negative Matrix Factorization," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 184-191.

[31] Liu J, Wang C, Gao J, et al. Multi-view clustering via joint nonnegative matrix factorization. Proceedings of the 2013 SIAM International Conference on Data Mining, Philadelphia: Society for Industrial and Applied Mathematics, 2013: 252-260.

[32] X. Zhang, L. Zhao, L. Zong, X. Liu and H. Yu, "Multi-view Clustering via Multi-manifold Regularized Nonnegative Matrix Factorization," 2014 IEEE International Conference on Data Mining, Shenzhen, China, 2014, pp. 1103-1108.

[33] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang. “More Diverse Means Better: Multimodal Deep Learning Meets Remote Sensing Imagery Classification, IEEE Transactions on Geoscience and Remote Sensing,” 2021, 59(5): 4340-4354.

[34] D. Hong, N. Yokoya, G. Xia, J. Chanussot, X. Zhu. " X-ModalNet: A Semi-Supervised Deep Cross-Modal Network for Classification of Remote Sensing Data," ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 167: 12-23.

[35] T. Xia, D. Tao, T. Mei and Y. Zhang, "Multiview Spectral Embedding," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 40, no. 6, pp. 1438-1446, Dec. 2010.

[36] Z. Deng, K. Choi, Y. Jiang and S. Wang, "Generalized Hidden-Mapping Ridge Regression, Knowledge-Leveraged Inductive Transfer Learning for Neural Networks, Fuzzy Systems and Kernel Methods," in IEEE Transactions on Cybernetics, vol. 44, no. 12, pp. 2585-2599, Dec. 2014.

[37] K. Xia et al., "TSK Fuzzy System for Multi-View Data Discovery Underlying Label Relaxation and Cross-Rule & Cross-View Sparsity Regularizations," in IEEE Transactions on Industrial Informatics, vol. 17, no. 5, pp. 3282-3291, May 2021.

[38] Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K. R., & Zien, A. (2009, December). Efficient and accurate lp-norm multiple kernel learning. In NIPS (Vol. 22, No. 22, pp. 997-1005).

[39] Cheng G, Han J. A survey on object detection in optical remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 117: 11-28.

[40] Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.

[41] Yang, J., & Zhang, Y. (2021). Home Textile Pattern Emotion Labeling Using Deep Multi-View Feature Learning. Frontiers in Psychology, 12.

[42] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint 2,1-norms minimization. In Neural Information Processing Systems (NIPS), Vancouver, Canada, December 2010.

[43] D. Hong, N. Yokoya, J. Chanussot, J. Xu, X. Zhu. Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 158, 35-49.

[44] D. Hong, N. Yokoya, J. Chanussot, J. Xu, X. Zhu. Joint and progressive subspace analysis (JPSA) with spatial-spectral manifold alignment for semisupervised hyperspectral dimensionality reduction. IEEE Transactions on Cybernetics, 2020, DOI: 10.1109/TCYB.2020.3028931.

[45] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, X. Zhu. "Invariant Attribute Profiles: A Spatial-Frequency Joint Feature Extractor for Hyperspectral Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no.6, pp. 3791-3808, 2020.