Tensor-based Multi-view Feature Selection with Applications to Brain Diseases

Bokai Cao*, Lifang He*‡, Xiangnan Kong†, Philip S. Yu*, Zhifeng Hao‡ and Ann B. Ragin§
*Department of Computer Science, University of Illinois at Chicago, IL, USA;
caobokai@uic.edu, lifanghescut@gmail.com, psyu@cs.uic.edu
†Department of Computer Science, Worcester Polytechnic Institute, MA, USA; xkong@wpi.edu
‡Department of Computer Science, Guangdong University of Technology, Guangzhou, China; zfhao@gdut.edu.cn
§Department of Radiology, Northwestern University, IL, USA; ann-ragin@northwestern.edu

***Abstract*—In the era of big data, we can easily access information from multiple views which may be obtained from different sources or feature subsets. Generally, different views provide complementary information for learning tasks. Thus, multi-view learning can facilitate the learning process and is prevalent in a wide range of application domains. For example, in medical science, measurements from a series of medical examinations are documented for each subject, including clinical, imaging, immunologic, serologic and cognitive measures which are obtained from multiple sources. Specifically, for brain diagnosis, we can have different quantitative analysis which can be seen as different feature subsets of a subject. It is desirable to combine all these features in an effective way for disease diagnosis. However, some measurements from less relevant medical examinations can introduce irrelevant information which can even be exaggerated after view combinations. Feature selection should therefore be incorporated in the process of multi-view learning. In this paper, we explore tensor product to bring different views together in a joint space, and present a dual method of tensor-based multi-view feature selection (DUAL-TMFS) based on the idea of support vector machine recursive feature elimination. Experiments conducted on datasets derived from neurological disorder demonstrate the features selected by our proposed method yield better classification performance and are relevant to disease diagnosis.**

***Index Terms*—tensor; brain diseases; multi-view learning; feature selection.**

I. Introduction

Many neurological disorders are characterized by ongoing injury that is clinically silent for prolonged periods and irreversible by the time symptoms first present. New approaches for detection of early changes in subclinical periods would afford powerful tools for aiding clinical diagnosis, clarifying underlying mechanisms and informing neuroprotective interventions to slow or reverse neural injury for a broad spectrum of brain disorders, including HIV infection on brain [10], [12], Alzheimer's disease [30], Parkinson's Disease, Schizophrenia, Depression, *etc.* Early diagnosis has the potential to greatly alleviate the burden of brain disorders and the ever increasing costs to families and society. For example, total healthcare costs for those 65 and older, are more that three times higher in those with Alzheimer's and other dementias [15].

As diagnosis of neurological disorder is extremely challenging, many different diagnosis tools and methods have been developed to obtain a large number of measurements from various examinations and laboratory tests. Information may be available for each subject for clinical, imaging, immunologic, serologic, cognitive and other parameters, as shown in Figure 1. In Magnetic Resonance Imaging (MRI) examination, for example, multiple strategies are used to interrogate the brain. Volumetric measurements of brain parenchymal and ventricular structures, and of major tissue classes (*e.g.* white matter, gray matter and CSF) can be derived. Volumetric measurements can also be quantified for a large number of individual brain regions and landmarks. While a single MRI examination can yield a vast amount of information concerning brain status at different levels of analysis, it is difficult to consider all available measures simultaneously, since they have different physical meanings and statistic properties. Capability for simultaneous consideration of measures coming from multiple groups is potentially transformative for investigating disease mechanisms and for informing therapeutic interventions.

As mentioned above, medical science witnesses everyday measurements from a series of medical examinations documented for each subject, including clinical, imaging, immunologic, serologic and cognitive measures. Each group of measures characterize the health state of a subject from different aspects. Conventionally this type of data is named as *multi-view data*, and each group of measures form a distinct *view* characterizing subjects in one specific feature space. An intuitive idea is to combine them to improve the learning performance, while simply concatenating features from all views and transforming a multi-view data into a single-view data would fail to leverage the underlying correlations between different views. We observe that tensors are higher order arrays that naturally generalize the notions of vectors and matrices to multiple dimensions. In this paper, we propose to use a tensor-based approach to model features (views) and their correlations hidden in the original multi-view data. Taking the tensor product of their respective feature spaces corresponds to the interaction of multiple views.

In the multi-view setting for neurological disorder, or for medical studies in general, however, a critical problem is that there may be limited subjects available yet introducing a large number of measurements. Within the multi-view data, not all features in different views are relevant to the learning task,

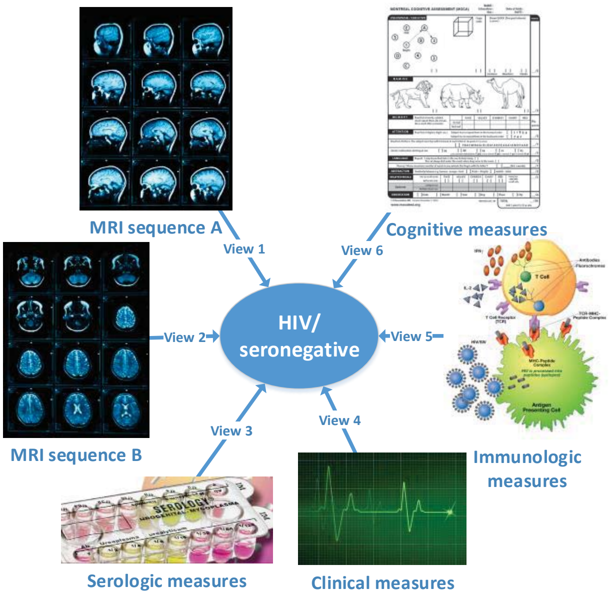

Fig. 1. An example of multi-view learning in medical studies.

and some irrelevant features may introduce unexpected noise. The irrelevant information can even be exaggerated after view combinations thereby degrading performance. Therefore, it is necessary to take care of feature selection in the learning process. Feature selection results can also be used by researchers to find biomarkers for brain diseases. Such biomarkers are clinically imperative for detecting injury to the brain in the earliest stage before it is irreversible. Valid biomarkers can be used to aid diagnosis, monitor disease progression and evaluate effects of intervention [13].

Considering feature selection, most of the existing studies can be categorized as filter models [17], [20] and embedded models based on sparsity regularization [7], [6], [26], [27]. While in this paper, we focus on wrapper models for feature selection. We propose a dual method of tensor-based multi-view feature selection (DUAL-TMFS), taking care of both the input space and the reconstructed tensor product space and exploiting their underlying correlations. In addition, our proposed method can naturally extend to many views and nonlinear kernels. Empirical studies on datasets collected from the Chicago Early HIV Infection Study [19] demonstrate that the proposed method can obtain better accuracy for classification tasks on multi-view feature selection than compared approaches. While the empirical studies are based on medical data from a clinical application in HIV infection on brain, the DUAL-TMFS technique developed for detecting brain anomalies have considerable promise for early diagnosis for other neurological disorders.

For the rest of the paper, we first state the problem of multi-view feature selection for classification and introduce related notations in section II. Then we introduce our DUAL-TMFS algorithm in section III. Experimental results are discussed in section IV. In section VI, we conclude the paper.

II. Problem Definition

In this section, we state the problem of multi-view feature selection for classification and introduce the notation. Table I lists some basic symbols that will be used throughout the paper. Note that although we use the same symbol to represent a set of data instances and the space that contains them, it is always clear from the context what we mean.

Suppose we have a multi-view classification task with n labeled instances represented from m different views: $\left\{\left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \cdots, \mathbf{x}_i^{(m)}, y_i\right)\right\}_{i=1}^{n}$, $\mathbf{x}_i^{(v)} \in \mathbb{R}^{I_v}$, $i \in \{1, \cdots, n\}$, $v \in \{1, \cdots, m\}$, where I_v is the dimensionality of the v-th view, and $y_i \in \{-1, 1\}$ is the class label of the i-th instance. We denote $\mathcal{X}_i = \{\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(m)}\}$, $\mathcal{X}^{(v)} = \{\mathbf{x}_1^{(v)}, \cdots, \mathbf{x}_n^{(v)}\}$, $\mathcal{Y} = \{y_1, \cdots, y_n\}$, and $\mathcal{D} = \{(\mathcal{X}_1, y_1), \cdots, (\mathcal{X}_n, y_n)\}$, respectively. The task of multi-view classification is to find a classifier function $f : \mathbb{R}^{I_1} \times \cdots \times \mathbb{R}^{I_m} \rightarrow \{-1, 1\}$ that correctly predicts the label of an unseen instance $\mathcal{X} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$.

One of the major challenges of multi-view classification comes from the fact that the combination of multiple views can potentially incur redundant and even conflicting information which is unfavorable for classifier learning. In order to tackle this problem, feature selection has been the focus of interest for quite some time and much work has been done in a supervised setting. A straightforward solution is to handle each view separately and conduct feature selection independently. This paradigm is based on the assumption that each view is sufficient on its own to learn the target concept [29]. However, individual views can often provide complementary information to each other leading to improved performance in real-world applications.

More generally, learning that involves conceptual *multi-view* is not just providing tools to analyze the data in multiple ways, which is more about managing the correlations among different views. Most previous feature selection approaches focus on exploiting multi-view features simultaneously to facilitate the learning process, which usually use the reconstructed data to represent the original multi-view information and perform analysis, such as the method (a) and method (b) shown in Figure 2. However, intrinsic properties of raw multi-view features and hidden relationships between the original data and its reconstruction are totally ignored in these methods.

Taking into account the latent interactions among views and the redundancy triggered by multiple views, in this paper, we aim at combining multiple features in a principled manner and performing feature selection to obtain a consensus and discriminative low-dimensional feature representation. In particular, we will leverage the relationship between the original multi-view features and reconstructed data to facilitate the implementation of feature selection.

III. Proposed Method

As noted in the introduction, one of the key issues for multi-view classification is to choose an appropriate tool to

TABLE I
LIST OF SYMBOLS

Symbol	Definition and Description
s	each lowercase letter represents a scale
$\mathbf{v}$	each boldface lowercase letter represents a vector
$\mathbf{M}$	each boldface capital letter represents a matrix
$\mathcal{T}$	each calligraphic letter represents a tensor, set or space
$\otimes$	denotes tensor product
$\langle .,. \rangle$	denotes inner product
$\lvert . \rvert$	denotes absolute value
$\lVert . \rVert_F$	denotes (Frobenius) norm of vector, matrix or tensor

model features (views) and their correlations hidden in the original multi-view features, since this directly determines how information will be used. The concept of tensor serves as a backbone for incorporating multi-view features into a consensus representation by means of tensor product, where the complex multiple relationships among views are embedded within the tensor structures. By mining structural information contained in the tensor, knowledge of multi-view features can be extracted and used to establish a predictive model. In this paper, we propose a dual method of tensor-based multi-view feature selection (DUAL-TMFS) in the tensor product space inspired by the idea of support vector machine recursive feature elimination (SVM-RFE) [9]. Our goal is to select useful features in conjunction with the classifier and simultaneously exploit the correlations among multiple views.

A. Tensor Propagation for Multiple Views

We start by introducing some related concepts and notation about tensors, and conceptually analyzing our motivation of utilizing tensor to organize all the multi-view information.

Tensors are higher order arrays that generalize the notions of vectors (first-order tensors) and matrices (second-order tensors), whose elements are indexed by more than two indices. Each index expresses a *mode* of variation of the data and corresponds to a coordinate direction. The number of variables in each mode indicates the dimensionality of a mode. The order of a tensor is determined by the number of its modes. The use of this data structure has been advocated in virtue of certain favorable properties. A key to this work is to borrow the tensor structure to fuse all possible dependence relationships among different views. We first recall the definition of tensor product (*i.e.*, outer product) of two vectors and then give a formal mathematical definition of the tensor, which provides an intuitive understanding of the algebraic structure of the tensor.

DEFINITION *1 (Tensor product):* The tensor product of two vectors $\mathbf{x} \in \mathbb{R}^{I_1}$ and $\mathbf{y} \in \mathbb{R}^{I_2}$, denoted by $\mathbf{x} \otimes \mathbf{y}$, represents a matrix with the elements $(\mathbf{x} \otimes \mathbf{y})_{i_1,i_2} = x_{i_1} y_{i_2}$.

DEFINITION *2 (Tensor):* A tensor is an element of the tensor product of vector spaces, each of which has its own coordinate system.

The tensor product of vector spaces forms an elegant algebraic structure for the theory of tensors. Such structure endows the tensor with the inherent advantage in representing real-world data, which naturally results from the interaction of multiple factors. Each mode of the tensor corresponds to one factor. For this reason, we conclude that the use of tensorial representation is a reasonable choice for adequately capturing the possible relationships among multiple views of data. Another advantage in representing all the multi-view information in the tensor data structure is that we can flexibly explore those useful knowledge in the tensor product space by virtue of tensor-based techniques.

Based on the definition of tensor product of two vectors, we can then express $\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z}$ as a third-order tensor in $\mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \mathbb{R}^{I_3}$, of which the elements are defined by $(\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z})_{i_1,i_2,i_3} = x_{i_1} y_{i_2} z_{i_3}$ for all values of the indices. Proceeding in the same way, $\mathcal{X} = (x_{i_1,\ldots,i_m})$ is used to denote an mth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ and its elements. For $v \in \{1, \cdots, m\}$, I_v is the dimensionality of $\mathcal{X}$ along the v-th mode. To indicate the object resulting by fixing the v-th mode index of $\mathcal{X}$ to be i_v, we introduce the generic subscript : and denote by $\mathcal{X}_{:,\ldots,:,i_v,:,\ldots,:}$.

In addition, we define the inner product and norm associated with tensor, which will be used in the following.

DEFINITION *3 (Inner product):* The inner product of two same-sized tensors $\mathcal{X}, \mathcal{Z} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ is defined as the sum of the products of their elements:

$$\langle \mathcal{X}, \mathcal{Z} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} x_{i_1,\ldots,i_m} z_{i_1,\ldots,i_m} \tag{1}$$

Most importantly, note that for tensors $\mathcal{X} = \mathbf{x}^{(1)} \otimes \cdots \otimes \mathbf{x}^{(m)}$ and $\mathcal{Z} = \mathbf{z}^{(1)} \otimes \cdots \otimes \mathbf{z}^{(m)}$, it holds that

$$\langle \mathcal{X}, \mathcal{Z} \rangle = \langle \mathbf{x}^{(1)}, \mathbf{z}^{(1)} \rangle \cdots \langle \mathbf{x}^{(m)}, \mathbf{z}^{(m)} \rangle \tag{2}$$

For the sake of brevity, in the following we will use the notation $\prod_{i=1}^{m} \otimes \mathbf{x}^{(i)}$ and $\prod_{i=1}^{m} \langle \mathbf{x}^{(i)}, \mathbf{z}^{(i)} \rangle$ to denote $\mathbf{x}^{(1)} \otimes \cdots \otimes \mathbf{x}^{(m)}$ and $\langle \mathbf{x}^{(1)}, \mathbf{z}^{(1)} \rangle \cdots \langle \mathbf{x}^{(m)}, \mathbf{z}^{(m)} \rangle$, respectively.

DEFINITION *4 (Tensor norm):* The norm of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ is defined to be the square root of the sum of all elements of the tensor squared, *i.e.*,

$$\lVert \mathcal{X} \rVert_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle} = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} x_{i_1,\ldots,i_m}^2} \tag{3}$$

As can be seen, the norm of a tensor is a straightforward generalization of the usual Frobenius norm for matrices and of the Euclidean or l_2 norm for vectors.

B. Multi-view SVM in the Tensor Setting

Following the introduction above to the concepts of tensors, we describe how multi-view classification can be consistently formulated and implemented in the framework of the standard SVM in the tensor setting.

By the reasoning given in section III-A, we use tensor product operation to bring m-view feature vectors of each instance together, leading to a tensorial representation for common structure across multiple views, and allowing us to adequately diffuse relationships and encode information among multi-view features. In this manner, we have essentially transformed the multi-view classification task from an independent domain of each view $\{(\mathcal{X}^{(1)}, \cdots, \mathcal{X}^{(m)}), \mathcal{Y}\}$ to a consensus domain $\{\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(m)}, \mathcal{Y}\}$ as a tensor classification problem.

For the sake of simplicity, we are slightly abusing notation by using $\mathcal{X}_i$ to denote $\prod_{v=1}^{m} \otimes \mathbf{x}_i^{(v)}$. Then the dataset of labeled multi-view instances can be represented as $\mathcal{D} = \{(\mathcal{X}_1, y_1), \cdots, (\mathcal{X}_n, y_n)\}$. Note that each multi-view instance $\mathcal{X}_i$ is an mth-order tensor that lies in the tensor product space $\mathbb{R}^{I_1 \times \cdots \times I_m}$, but one must keep in mind that each element of $\mathcal{X}_i$ is the tensor product of multi-view features in the input space, which we denote by $x_{i(i_1,\ldots,i_m)}$. Now, based on the definitions of inner product and tensor norm, we can formulate multi-view classification as a global convex optimization problem in the framework of the standard SVM as follows:

$$\min_{\mathcal{W},b,\xi} \frac{1}{2} \|\mathcal{W}\|_F^2 + C \sum_{i=1}^{n} \xi_i \tag{4}$$

$$\text{s.t. } y_i(\langle \mathcal{W}, \mathcal{X}_i \rangle + b) \geq 1 - \xi_i \tag{5}$$

$$\xi_i \geq 0, \forall i = 1, \cdots, n. \tag{6}$$

where $\mathcal{W}$ can be regarded as the weight tensor of the separating hyperplane in the tensor product space $\mathbb{R}^{I_1 \times \cdots \times I_m}$, b is the bias, ξ_i is the error of the i-th training sample, and C is the trade-off between the margin and empirical loss. As such it can be solved with the use of optimization techniques developed for SVM, and the weight tensor of $\mathcal{W}$ can be obtained from

$$\mathcal{W} = \sum_{i=1}^{n} \alpha_i y_i \mathcal{X}_i \tag{7}$$

where α_i is the dual variable corresponding to each instance. The resulting decision function is

$$f(\mathcal{X}) = \text{sign}(\langle \mathcal{W}, \mathcal{X} \rangle + b) \tag{8}$$

where $\mathcal{X}$ denotes a test multi-view instance given by the tensor product of its multi-view features $\mathbf{x}^{(v)}$ for all $v \in \{1, \cdots, m\}$. We simply call the model as *multi-view SVM*.

Despite this property, there are two major drawbacks incurred by the combination of multiple views. First, the dimensionality of the resulting tensor in a multi-view dataset can be extremely large, which grows at an exponential rate with respect to the number of views. Direct application of the *multi-view SVM* will suffer from the curse of dimensionality. Second, such tensors may contain much redundant and irrelevant information due to the intrinsic multi-view property, which will negatively influence the performance of the learning process.

Therefore, in order to implement multi-view classification using *multi-view SVM*, it is necessary to perform dimensionality reduction by feature extraction or selection to concentrate multi-view information and improve tensorial representation. Many tensor-based algorithms have been proposed as dimensionality reduction for classification problems. However, to the best of our knowledge, all of them discard the original multi-view features after constructing tensors. In the following, we investigate their relationship to each other and proceed to develop a wrapper feature selection approach DUAL-TMFS.

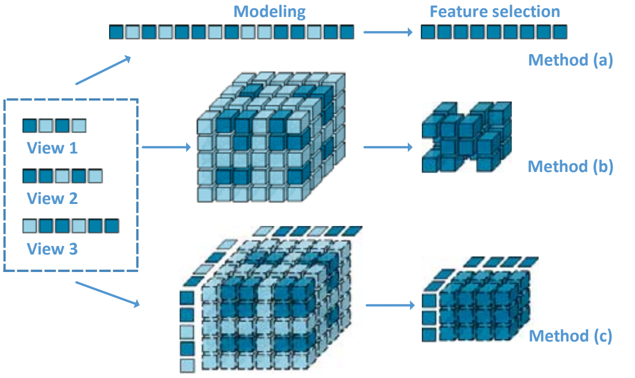

Fig. 2. Schematic view of the key differences among three strategies of multi-view feature selection. Method (a) concatenates features from all views in the input space. Method (b) converts multiple views into a tensor and directly performs feature selection in the tensor product space. Our method (c) efficiently conducts feature selection in the input space while effectively leveraging relationships between the original data and its reconstruction in the tensor product space.

C. Dual Feature Selection in the Tensor Product Space

Based on the *multi-view SVM* classifier in the tensor setting, in this subsection, we approach the problem of identifying and concentrating multi-view knowledge via tensors by proposing the linear DUAL-TMFS method. We will extend it to the nonlinear case in the next subsection.

Inspired by SVM-RFE [9], we can see from Eq. (8) that the inner product of weight tensor $\mathcal{W} = (w_{i_1,\ldots,i_m})$ and input tensor $\mathcal{X} = (x_{i_1,\ldots,i_m})$ determines the value of $f(\mathcal{X})$. Intuitively, the input features that are weighted by the largest absolute values influence most on the classification decision, and correspond to the most informative features. Therefore, the absolute weights $|w_{i_1,\ldots,i_m}|$ or the square of the weights $(w_{i_1,\ldots,i_m})^2$ can be used as feature ranking criterion to select the most discriminative feature subset. Based on this observation, we can conduct feature selection on *multi-view SVM*.

Let us denote the ranking score of each feature $x_{i_1,\ldots,i_m}$ as $r_{i_1,\ldots,i_m}$. Our target is to perform feature elimination in the tensor product feature space by

$$\underset{i_1,\cdots,i_m}{\text{argmin}} (r_{i_1,\ldots,i_m}) \tag{9}$$

SVM-RFE performs SVM-based feature selection in the vector space, as the method (a) shown in Figure 2. A straightforward approach, which can be seen as a natural tensorial extension of SVM-RFE, is to directly perform feature elimination in the tensor product space using the following feature ranking criterion:

$$r_{i_1,\ldots,i_m} = (w_{i_1,\ldots,i_m})^2 \tag{10}$$

As the method (b) shown in Figure 2, however, the number of variables $w_{i_1,\ldots,i_m}$ is equivalent to the dimensionality of the resulting tensors in tensor product space. Obviously, it is computationally intractable to enumerate all values of $w_{i_1,\ldots,i_m}$ in such a high-dimensional tensor product space. On the other hand, the original multi-view features usually contain much redundant and irrelevant features. It can be further exaggerated over the manipulation of tensor product, thereby degrading the generalization performance. In order to overcome these problems, it would be desirable to remove irrelevant features before manipulating the tensor product.

Considering that each view has specific statistical properties and its intrinsic physical meanings, we conduct multi-view feature selection in the input space and maintain independent rankings of features in each view. We leverage the weight coefficients $\mathcal{W}$ in the tensor product space to facilitate the implementation of feature selection in the input space. That is, for the v-th view, supposing $\mathbf{x}^{(v)} = [x_1^{(v)}, \cdots, x_{I_v}^{(v)}]$, the ranking score of the feature $x_{i_v}^{(v)}, i_v \in \{1, \cdots, I_v\}$ in the input space is $r_{i_v}^{(v)}$, which means $r_{i_v}^{(v)}$ is a function of $w_{i_1,\ldots,i_m}$.

Now we can formulate the problem in terms of the process, for which we need to minimize the following function in each view $v \in \{1, \cdots, m\}$:

$$\underset{i_v}{\operatorname{argmin}} \left(r_{i_v}^{(v)} \right) \tag{11}$$

An alternative approach is to evaluate the value of $r_{i_v}^{(v)}$ from $w_{i_1,\ldots,i_m}$ by virtue of the relationship between the input space and the tensor product space. Based on the definition of the tensor product, we can see that the feature $x_{i_v}^{(v)}$ in the input space will diffuse to $\mathcal{X}_{:,\ldots,:,i_v,:,\ldots,:}$ in the tensor product space, thus to $\mathcal{W}_{:,\ldots,:,i_v,:,\ldots,:}$. Intuitively, it means that the contribution of $x_{i_v}^{(v)}$ determining the value of decision function $f(\mathcal{X})$ transfers to $\mathcal{X}_{:,\ldots,:,i_v,:,\ldots,:}$. For this reason, the ranking score of $x_{i_v}^{(v)}$ can be estimated from the elements of $\mathcal{W}_{:,\ldots,:,i_v,:,\ldots,:}$. To realize such purpose, we set $r_{i_v}^{(v)}$ equal to the sum of the square of all elements in $\mathcal{W}_{:,\ldots,:,i_v,:,\ldots,:}$, which is given as follows:

$$r_{i_v}^{(v)} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_{v-1}=1}^{I_{v-1}} \sum_{i_{v+1}=1}^{I_{v+1}} \cdots \sum_{i_m=1}^{I_m} (w_{i_1,\cdots,i_m})^2 \tag{12}$$

By substituting the exact solution given in Eq. (7) into the right-hand side of this equality, we find that

$$r_{i_v}^{(v)} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_{i(i_v)}^{(v)} x_{j(i_v)}^{(v)} \prod_{1 \le k \le m}^{k \ne v} \langle \mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)} \rangle \tag{13}$$

In this way, compared with performing feature selection in the tensor product space, the computational complexity is largely reduced, since irrelevant and redundant features can be detected by the classifier constructed in the tensor product space, but removed in the input space, which concentrates the multi-view information within tensor operations. Conducting feature selection in the input space is superior in terms of better readability and interpretability, because it maintains the physical meanings of the original features without any manipulation. This property has its significance in many real-world applications such as finding clinical markers to a specific disease.

Nevertheless, although this is expected to improve tensorial representation of multi-view data and perform feature selection for multi-view classification, it can result in potential over-fitting, since the number of variables $w_{i_1,\ldots,i_m}$ grows at an exponential rate as m (*i.e.*, the number of views) increases. Especially in medical studies, there may be limited subjects available yet introducing a large number of measurements in many views. Therefore, the problem reduces to improving the generalization capability of *multi-view SVM* in the tensor setting, for which we need a more sophisticated approach to reduce the number of variable $w_{i_1,\ldots,i_m}$ (*i.e.*, the number of elements of $\mathcal{W}$ that need to be estimated) and facilitate feature selection without incurring extensive computation.

In the context of supervised tensor learning, tensor decompositions are usually used to reduce the number of unknown tensors (*i.e.*, the dimensionality of tensor), and meanwhile avoid overfitting. Following assumptions in the supervised tensor learning framework [24], here we assume that $\mathcal{W}$ can be decomposed as $\mathcal{W} = \prod_{v=1}^{m} \otimes \mathbf{w}^{(v)}$, where $\mathbf{w}^{(v)} = [w_1^{(v)}, \cdots, w_{I_v}^{(v)}]$. By applying Eqs. (2) and (3), we can then represent the optimization problem in Eqs. (4)-(6) as:

$$\min_{\mathbf{w}^{(v)}, b, \xi} \frac{1}{2} \prod_{v=1}^{m} \left\| \mathbf{w}^{(v)} \right\|_F^2 + C \sum_{i=1}^{n} \xi_i \tag{14}$$

$$\text{s.t. } y_i \left(\prod_{v=1}^{m} \left\langle \mathbf{w}^{(v)}, \mathbf{x}_i^{(v)} \right\rangle + b \right) \ge 1 - \xi_i \tag{15}$$

$$\xi_i \ge 0, \forall i = 1, \cdots, n. \tag{16}$$

thus the optimal decision function is:

$$f(\mathcal{X}) = \text{sign} \left(\prod_{v=1}^{m} \left\langle \mathbf{w}^{(v)}, \mathbf{x}^{(v)} \right\rangle + b \right) \tag{17}$$

Clearly, in this manner, the number of variables with respect to $\mathcal{W}$ is greatly reduced from $\prod_{v=1}^{m} I_v$ to $\sum_{v=1}^{m} I_v$. Moreover, from Eq. (17), we can see that the influence of input feature $x_{i_v}^{(v)}$ on the value of decision function $f(\mathcal{X})$ constructed in the tensor product space is determined only by its corresponding weight coefficient $w_{i_v}^{(v)}$, *i.e.*, the feature ranking criterion defined in Eq. (12) can be simplified as:

$$r_{i_v}^{(v)} = \left(w_{i_v}^{(v)} \right)^2 \tag{18}$$

THEOREM *1: The ranking criterion Eq. (18) is equivalent to Eq. (12) for each view.*

Proof. Based on the definition of tensor product, we have $w_{i_1,\ldots,i_m} = w_{i_1}^{(1)} \cdots w_{i_m}^{(m)}$. Substituting this into Eq. (12), it can be written as:

$$\begin{aligned} r^{(v)}_{i_v} =& \sum_{i_1}\cdots\sum_{i_{v-1}}\sum_{i_{v+1}}\cdots\sum_{i_m}(w_{i_1,\cdots,i_m})^2 \\ =& \sum_{i_1}\cdots\sum_{i_{v-1}}\sum_{i_{v+1}}\cdots\sum_{i_m}\left(w^{(1)}_{i_1}\cdots w^{(m)}_{i_m}\right)^2 \\ =& \left(w^{(v)}_{i_v}\right)^2 \prod_{1\le j\le m}^{j\neq v}\left\|\mathbf{w}^{(j)}\right\|_F^2 \\ =& P^{(-v)}\left(w^{(v)}_{i_v}\right)^2 \end{aligned} \tag{19}$$

where $P^{(-v)} = \prod_{1\le j\le m}^{j\neq v}\|\mathbf{w}^{(j)}\|_F^2$. For the v-th mode, the multiplier $P^{(-v)}$ is constant and non-negative, thus has no effect on ranking orders. The proof is complete.

Now we illustrate how to solve the optimization problem in Eqs. (14)-(16). In an iterative manner, we can update the variables associated with a single mode at each iteration. That is, for the v-th mode, we need to fix variables in other modes and solve the following optimization problem:

$$\min_{\mathbf{w}^{(v)},b^{(v)},\xi^{(v)}} \frac{P^{(-v)}}{2}\left\|\mathbf{w}^{(v)}\right\|_F^2 + C\sum_{i=1}^{n}\xi_i^{(v)} \tag{20}$$

$$\text{s.t. } y_i\left(Q_i^{(-v)}\left\langle\mathbf{w}^{(v)},\mathbf{x}_i^{(v)}\right\rangle + b^{(v)}\right) \ge 1-\xi_i^{(v)} \tag{21}$$

$$\xi_i^{(v)} \ge 0, \forall i = 1,\cdots,n. \tag{22}$$

where $P^{(-v)}$ and $Q_i^{(-v)}$ are constants that denote $P^{(-v)} = \prod_{1\le j\le m}^{j\neq v}\|\mathbf{w}^{(j)}\|_F^2$ and $Q_i^{(-v)} = \prod_{1\le j\le m}^{j\neq v}\langle\mathbf{w}^{(j)},\mathbf{x}_i^{(j)}\rangle$.

Let $\mathbf{x}_i^{(v)'} = (Q_i^{(-v)}/\sqrt{P^{(-v)}})\mathbf{x}_i^{(v)}$ and $\mathbf{w}^{(v)'} = \sqrt{P^{(-v)}}\mathbf{w}^{(v)}$, then the optimization problem in Eqs. (20)-(22) is equivalent to the following problem:

$$\min_{\mathbf{w}^{(v)'},b^{(v)},\xi^{(v)}} \frac{1}{2}\left\|\mathbf{w}^{(v)'}\right\|_F^2 + C\sum_{i=1}^{n}\xi_i^{(v)} \tag{23}$$

$$\text{s.t. } y_i\left(\left\langle\mathbf{w}^{(v)'},\mathbf{x}_i^{(v)'}\right\rangle + b^{(v)}\right) \ge 1-\xi_i^{(v)} \tag{24}$$

$$\xi_i^{(v)} \ge 0, \forall i = 1,\cdots,n. \tag{25}$$

which reduces to the standard linear SVM, and thus can be efficiently solved by available algorithms, obtaining $\mathbf{w}^{(v)}$ as follows:

$$\mathbf{w}^{(v)} = \frac{1}{P^{(-v)}}\sum_{i=1}^{n}Q_i^{(-v)}\alpha_i^{(v)}y_i\mathbf{x}_i^{(v)} \tag{26}$$

where $\alpha_i^{(v)}$ is the dual variable corresponding to each instance in the v-th mode, obtained in Eqs. (23)-(25).

It is illustrated in Figure 2 that, the method (c) leveraging the ranking criterion Eq. (18) jointly considers the input space and the tensor product space, and effectively exploits their underlying relationship. We summarize our proposed dual method of multi-view feature selection (DUAL-TMFS) in Figure 3.

```
Input:
  - Training examples in multiple views:
      $\mathfrak{X}^{(v)} = \{\mathbf{x}_1^{(v)},\cdots,\mathbf{x}_n^{(v)}\}, v = 1,2,\cdots,m$
  - Class labels: $\mathfrak{Y} = \{y_1,\cdots,y_n\}$
  - Number of features selected in each view: $p_v$
Initialize:
  - Subset of surviving features: $\mathbf{s}^{(v)} = [1,2,\cdots,I_v]$
Iterate through each view $v$:
Repeat until length($\mathbf{s}^{(v)}$) $\le p_v$
  - Restrict training examples to good feature indices:
      $\mathfrak{X}^{(v)*} = \mathfrak{X}^{(v)}(\mathbf{s}^{(v)},:)$
  - Train the classifier: $\alpha = \text{SVM-TRAIN}(\mathfrak{X}^{(v)*},\mathfrak{Y})$
  - Compute the weight vector $\mathbf{w}^{(v)}$ according to Eq. (26)
  - Compute the ranking criteria $\mathbf{r}^{(v)}$ according to Eq. (18)
  - Find the feature with smallest ranking criterion:
      $f = \arg\min(\mathbf{r}^{(v)})$
  - Eliminate the feature with smallest ranking criterion:
      $\mathbf{s}^{(v)} = \mathbf{s}^{(v)}(1{:}f\text{-}1, f\text{+}1{:}\text{length}(\mathbf{s}^{(v)}))$
Output:
  - Subset of selected features in each view:
      $\mathbf{s}^{(v)}, v = 1,2,\cdots,m$
```

Fig. 3. The DUAL-TMFS algorithm

D. Extension to Nonlinear Kernels

As discussed above, tensor is an effective approach of capturing correlations across multiple views. However, correlations between features within the same view are not considered by taking the tensor product of features in different views. In such case, we can replace the linear kernel with a nonlinear kernel. Through implicitly projecting features into a high dimensional space within each view, a nonlinear kernel can work together with the tensor tools to exploit correlations across different views as well as those within each view.

In the case of nonlinear SVMs, we first represent optimization problem in Eqs. (23)-(25) in the dual form as:

$$\min_{\alpha}\frac{1}{2}\alpha^{(v)T}H\alpha^{(v)} - \alpha^{(v)T}\mathbf{1} \tag{27}$$

$$\text{s.t. } \sum_{i=1}^{n}\alpha_i^{(v)}y_i = 0 \tag{28}$$

$$0 \le \alpha_i^{(v)} \le C, \forall i = 1,\cdots,n. \tag{29}$$

where H is the matrix with elements $y_h y_k \kappa(\mathbf{x}_h^{(v)'},\mathbf{x}_k^{(v)'})$.

To compute the change in cost function caused by removing input component i_v in the v-th mode, one leaves the α's unchanged and one re-computes matrix H. This corresponds to computing $\kappa(\mathbf{x}_h^{(v)'}(-i_v),\mathbf{x}_k^{(v)'}(-i_v))$, yielding matrix $H(-i_v)$, where the notation $(-i_v)$ means that component i_v has been removed in the v-th mode. Thus, the feature ranking criterion for nonlinear SVMs is:

$$r_{i_v}^{(v)} = \frac{1}{2}(\alpha^{(v)T}H\alpha^{(v)} - \alpha^{(v)T}H(-i_v)\alpha^{(v)}) \tag{30}$$

The input corresponding to the smallest difference $r_{i_v}^{(v)}$ shall be removed. In the linear case, $\kappa(\mathbf{x}_h^{(v)'},\mathbf{x}_k^{(v)'}) = \langle\mathbf{x}_h^{(v)'},\mathbf{x}_k^{(v)'}\rangle$

and $\alpha^{(v)T} H \alpha^{(v)} = \|\mathbf{w}^{(v)'}\|_F^2$. Therefore $r_{i_v}^{(v)} = \frac{P^{(-v)}}{2}(w_{v_i}^{(v)})^2$, which is equivalent to the one we proposed in the previous section for linear SVMs.

IV. Experiments

In this section, we conduct experiments on datasets collected from HIV infected brain disease, to evaluate our proposed method in different aspects. In section IV-C, we have seven methods compared on the classification tasks composing of two views. Experiments extend to more than two views in section IV-D. In nonlinear cases, our method can still be effectively applied, as shown in section IV-E.

A. Data Collections

In order to evaluate the performance of multi-view feature selection for classification, we compare methods on datasets collected from the Chicago Early HIV Infection Study [19], which have 56 HIV and 21 seronegative control subjects enrolled. For each subject, hundreds of clinical, imaging, immunologic, serologic and cognitive measures were documented. This illustrates the curse of dimensionality because there are far more variables of interest than available subjects. Thus, it is important to incorporate feature selection in the learning process for disease diagnosis.

There are seven groups of measurements investigated in our datasets, including *neuropsychological tests*, *flow cytometry*, *plasma luminex*, *freesurfer*, *overall brain microstructure*, *localized brain microstructure*, *brain volumetry*. Each group can be regarded as a distinct view that partially reflects subject status, and measurements from different medical examinations can provide complementary information. Simultaneous consideration of all the data, exploiting correlations among multiple measurements can be transformative for investigating disease mechanisms and for informing therapeutic interventions. Different views are sampled to form multiple combinations. The datasets used for our experiments are summarized in Table II. Additionally, features are normalized within $[0, 1]$.

B. Compared Methods

In order to demonstrate the effectiveness of our multi-kernel learning approach, we compare the following methods:

- CF refers to single-kernel SVM applying on concatenated features.
- TPF refers to single-kernel SVM applying on the tensor product feature space [11]. By taking the tensor product of features from different views, adequate correlations among different views are exploited.
- LINEAR-MKL is a conventional multi-kernel method [5]. Different kernels can naturally correspond to different views. Through an optimization framework, weights can be learned that reflect the relative importance of different views. It implements a linear combination of multiple kernels.
- RFE-CF denotes the method that directly applies SVM-RFE on the concatenation of all the features [9].
- RFE-TPF denotes the method that SVM-RFE is applied on the tensor product feature space [21].
- MIQP-TPFS refers to the method of iterative tensor product feature selection with mixed-integer quadratic programming [21]. It explicitly considers the cross-domain interactions between two views in the tensor product feature space. The bipartite feature selection problem is formulated as an integer quadratic programming problem. A subset of features is selected that maximizes the sum over the submatrix of the original weight matrix.
- DUAL-TMFS is the proposed dual method of tensor-based multi-view feature selection in the tensor product feature space. It effectively exploits the correlations among different views in the tensor product feature space, and also efficiently completes feature selection in the input space at the same time.

A detailed comparison between these methods is summarized in Table III, against four dimensions on whether the schemes can conduct feature selection, discriminate against different views, be applicable to many views and compatible with nonlinear kernels. Note that sparsity regularization models [8], [27] are not considered as we focus on wrapper models in this paper, without looking into embedded models.

For a fair comparison, we use LibSVM [3] with linear kernel as the base classifier for all the compared methods. In the experiments, 3-fold cross validations are performed on balanced datasets. The soft margin parameter C is selected through a validation set. For all the methods with feature selection, the number of features selected is explicitly set to 50%.

C. Two Views

We first study the effectiveness of our proposed method on the task of learning from two views. Results on D2.1 and D2.2 are shown in Table IV, where the average performances of the compared methods with standard deviations are reported with respect to four evaluation metrics: accuracy, precision, recall and F1-measure.

In comparison of the top three methods not conducting feature selection, there is no clear advantage for any of the methods. Performance can vary depending on datasets, if the redundancy coming from different views is not taken care of. Thus, it is necessary to select discriminative features and eliminate redundant ones when multiple views are combined.

While considering feature selection, DUAL-TMFS significantly improves the accuracy over other methods by effectively pruning redundant and irrelevant features. On the other hand, simply applying SVM-RFE method on either the input space (*i.e.*, RFE-CF) or the tensor product feature space (*i.e.*, RFE-TPF) cannot achieve better performance. For RFE-CF, correlations between multiple views are not exploited when selecting features; while for RFE-TPF, features are directly selected in the tensor product space, leaving the potential of overfitting. MIQP-TPFS can take advantage of feature selection by maximizing the sum over the weight submatrix in the tensor product feature space.

TABLE II
SUMMARY OF DATASETS. "■" INDICATES THE VIEW IS SELECTED IN THE DATASET, WHILE "□" INDICATES NOT SELECTED. EACH NUMBER IN BRACES INDICATES THE NUMBER OF FEATURES IN EACH VIEW.

Name	D2.1	D2.2	D3.1	D3.2	D4.1	D4.2	D5.1	D5.2	D6.1	D6.2
#Views	2	2	3	3	4	4	5	5	6	6
neuropsychological tests (36)	□	□	■	□	□	□	■	□	■	■
flow cytometry (65)	□	□	□	□	■	■	■	■	■	■
plasma luminex (45)	■	□	■	■	■	□	□	■	■	■
freesurfer (28)	□	■	■	□	□	■	□	■	□	■
overall brain microstructure (21)	■	■	□	□	■	■	■	□	■	■
localized brain microstructure (54)	□	□	□	■	□	■	■	■	■	■
brain volumetry (12)	□	□	□	■	■	□	■	■	■	□

TABLE III
SUMMARY OF COMPARED METHODS.

Property	CF	TPF[11]	LINEAR-MKL[5]	RFE-CF[9]	RFE-TPF[21]	MIQP-TPFS[21]	DUAL-TMFS
conducting feature selection	×	×	×	√	√	√	√
discriminating different views	×	√	√	×	√	√	√
applicability to many views	√	√	√	√	×	×	√
compatibility with nonlinear kernels	√	×	×	√	×	×	√

D. Many Views

In real-world applications, there are usually more than two views. It is desirable to leverage all of them simultaneously. However, RFE-TPF and MIQP-TPFS need to explicitly compute the tensor product feature space, resulting in complexity and space complexity exponential to the number of views. They are therefore no longer feasible in the case of many views, due to high dimensionality of the tensor product feature space. Although DUAL-TMFS also exploits the correlations among different views in the tensor product feature space, it can efficiently complete feature selection in the input space. Thus, our proposed method have time complexity and space complexity linear with respect to the number of views and can naturally extend to more than two views. The experimental results are summarized at D3.1-D6.2 in Table IV.

As can be seen, neither CF nor RFE-CF performs well. This shows that simply concatenating all features across multiple views does not work well. We next consider schemes that discriminate different views. TPF performs badly as it computes the tensor product feature space, introducing some potentially irrelevant features coming from the correlations among multiple views. In general, LINEAR-MKL performs well in most cases by linearly weighting multiple kernels. However, by further performing feature selection, DUAL-TMFS achieves a significant improvement over other methods and always ranks first on F1 and accuracy, indicating that compared with approaches not distinguishing different views or not conducting feature selection, a better subset of discriminative features can be selected for classification by considering the correlations across multiple views based on tensor.

E. Nonlinear Kernels

As discussed above, tensor is an effective approach of capturing correlations across multiple views. However, correlations between features within the same view are not considered by taking the tensor product of features in different views. In such case, we can replace a linear kernel with a nonlinear kernel. Through implicitly projecting features into a high dimensional space within each view, a nonlinear kernel can work together with the tensor tools to exploit correlations across different views as well as that within each view.

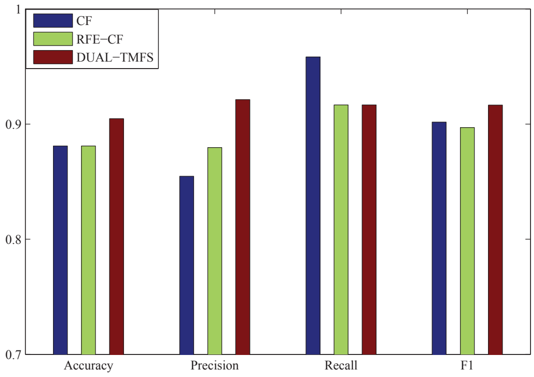

Fig. 4. Classification performance in the nonlinear case.

Here we replace the linear kernel with the RBF kernel for all the compared methods, and show experimental results in Figure 4. LINEAR-MKL is not applicable in the nonlinear case. Neither do TPF, RFE-TPF and MIQP-TPFS, because they need to explicitly compute the high dimensional feature space which is intractable when we apply a nonlinear kernel. It illustrates that DUAL-TMFS still outperforms other methods in the nonlinear case, in the sense of accuracy and F1-measure.

TABLE IV
CLASSIFICATION PERFORMANCE “AVERAGE SCORE (RANK)” IN THE LINEAR CASE. FOR EACH DATASET, THE TOP-3 METHODS ARE WITHOUT FEATURE SELECTION.

Datasets	Methods	Evaluations			
		Accuracy	Precision	Recall	F1
D2.1	CF	0.738 (2)	0.747 (2)	0.833 (1)	0.782 (2)
	TPF	0.595 (7)	0.623 (7)	0.833 (1)	0.698 (7)
	LINEAR-MKL	0.643 (5)	0.710 (4)	0.792 (4)	0.730 (5)
	RFE-CF	0.690 (3)	0.730 (3)	0.750 (7)	0.736 (4)
	RFE-TPF	0.643 (5)	0.657 (6)	0.792 (4)	0.713 (6)
	MIQP-TPFS	0.667 (4)	0.667 (5)	0.833 (1)	0.737 (3)
	DUAL-TMFS	0.762 (1)	0.784 (1)	0.792 (4)	0.786 (1)
D2.2	CF	0.548 (6)	0.657 (3)	0.500 (6)	0.553 (6)
	TPF	0.619 (3)	0.639 (5)	0.750 (2)	0.683 (4)
	LINEAR-MKL	0.667 (2)	0.730 (2)	0.667 (4)	0.692 (2)
	RFE-CF	0.524 (7)	0.607 (7)	0.500 (6)	0.540 (7)
	RFE-TPF	0.619 (3)	0.656 (4)	0.750 (2)	0.690 (3)
	MIQP-TPFS	0.595 (5)	0.638 (6)	0.667 (4)	0.648 (5)
	DUAL-TMFS	0.714 (1)	0.769 (1)	0.792 (1)	0.759 (1)
D3.1	CF	0.762 (2)	0.795 (2)	0.792 (2)	0.791 (2)
	TPF	0.690 (5)	0.692 (5)	0.833 (1)	0.753 (5)
	LINEAR-MKL	0.738 (3)	0.783 (3)	0.750 (5)	0.763 (3)
	RFE-CF	0.714 (4)	0.741 (4)	0.792 (2)	0.761 (4)
	DUAL-TMFS	0.833 (1)	0.926 (1)	0.792 (2)	0.846 (1)
D3.2	CF	0.690 (4)	0.727 (3)	0.750 (3)	0.734 (3)
	TPF	0.714 (2)	0.711 (4)	0.833 (2)	0.767 (2)
	LINEAR-MKL	0.714 (2)	0.822 (1)	0.667 (5)	0.709 (5)
	RFE-CF	0.667 (5)	0.692 (5)	0.750 (3)	0.718 (4)
	DUAL-TMFS	0.810 (1)	0.820 (2)	0.875 (1)	0.839 (1)
D4.1	CF	0.857 (4)	0.847 (4)	0.917 (3)	0.880 (4)
	TPF	0.833 (5)	0.838 (5)	0.875 (5)	0.855 (5)
	LINEAR-MKL	0.905 (2)	0.917 (2)	0.917 (3)	0.917 (2)
	RFE-CF	0.881 (3)	0.852 (3)	0.958 (1)	0.902 (3)
	DUAL-TMFS	0.929 (1)	0.926 (1)	0.958 (1)	0.939 (1)
D4.2	CF	0.857 (3)	0.886 (3)	0.875 (4)	0.874 (3)
	TPF	0.810 (5)	0.792 (5)	0.917 (1)	0.847 (5)
	LINEAR-MKL	0.905 (2)	0.917 (2)	0.917 (1)	0.917 (2)
	RFE-CF	0.833 (4)	0.878 (4)	0.833 (5)	0.852 (4)
	DUAL-TMFS	0.929 (1)	0.958 (1)	0.917 (1)	0.936 (1)
D5.1	CF	0.905 (2)	0.963 (1)	0.875 (3)	0.911 (3)
	TPF	0.810 (5)	0.812 (5)	0.875 (3)	0.837 (5)
	LINEAR-MKL	0.905 (2)	0.917 (4)	0.917 (2)	0.917 (2)
	RFE-CF	0.905 (2)	0.963 (1)	0.875 (3)	0.911 (3)
	DUAL-TMFS	0.952 (1)	0.963 (1)	0.958 (1)	0.958 (1)
D5.2	CF	0.881 (3)	0.915 (3)	0.875 (4)	0.892 (3)
	TPF	0.714 (5)	0.719 (5)	0.833 (5)	0.771 (5)
	LINEAR-MKL	0.905 (1)	0.917 (1)	0.917 (1)	0.917 (1)
	RFE-CF	0.857 (4)	0.847 (4)	0.917 (1)	0.880 (4)
	DUAL-TMFS	0.905 (1)	0.917 (1)	0.917 (1)	0.917 (1)
D6.1	CF	0.881 (4)	0.915 (4)	0.875 (3)	0.892 (4)
	TPF	0.833 (5)	0.838 (5)	0.875 (3)	0.855 (5)
	LINEAR-MKL	0.905 (2)	0.917 (3)	0.917 (1)	0.917 (2)
	RFE-CF	0.905 (2)	0.952 (2)	0.875 (3)	0.911 (3)
	DUAL-TMFS	0.952 (1)	1.000 (1)	0.917 (1)	0.956 (1)
D6.2	CF	0.905 (2)	0.921 (3)	0.917 (1)	0.917 (2)
	TPF	0.810 (5)	0.810 (5)	0.875 (3)	0.841 (5)
	LINEAR-MKL	0.905 (2)	0.917 (4)	0.917 (1)	0.917 (2)
	RFE-CF	0.905 (2)	0.952 (2)	0.875 (3)	0.911 (4)
	DUAL-TMFS	0.929 (1)	1.000 (1)	0.875 (3)	0.930 (1)

TABLE V
TOP-3 MEASURES SELECTED IN EACH VIEW.

neuropsychological tests : *Karnofsky Performance Scale, NART FSIQ, Rey Trial*
flow cytometry : *Tcells 4+8-, 3+56-16+NKT Cells 4+8-, Lymphocytes*
plasma luminex : *MMP-2, GRO, TGFa*
freesurfer : *Cerebral Cortex, Thalamus Proper, CC_Mid_Posterior*
overall brain microstructure : *MTR-CC, MTR-Hippocampus, MD-Cerebral-White-Matter*
localized brain microstructure : *MTR-CC_Mid_Anterior, FA-CC_Anterior, MTR-CC_Central*
brain volumetry : *Norm Peripheral Gray Volume, BPV, Norm Brain Volume*

F. Feature Evaluation

Table V lists the most discriminative measures selected by DUAL-TMFS. Our results are validated by literature on brain diseases. The Karnofsky Performance Status is the most widely used health status measure in HIV medicine and research [16]. [2] observes CD4+ T cell depletion during all stages of HIV disease. Mycoplasma membrane protein (MMP) is identified as a possible cofactor responsible for the progression of AIDS. The fronto-orbital cortex, one of the cerebral cortical areas, is mainly damaged in AIDS brains [28]. Whole brain MTR is reduced in HIV-1-infected patients [18]. [1] concludes HIV dementia is associated with specific gray matter volume reduction, as well as with generalized volume reduction of white matter.

V. RELATED WORK

Currently, representative methods for multi-view learning can be categorized into three groups [29]: co-training, multiple kernel learning, and subspace learning. Generally, the co-training style algorithm is a classic approach for semi-supervised learning, which trains alternatively to maximize the mutual agreement on different views. Multiple kernel learning algorithms combine kernels that naturally correspond to different views, either linearly [14] or nonlinearly [25], [4] to improve learning performance. Subspace learning algorithms learn a latent subspace, from which multiple views are generated. Multiple kernel learning and subspace learning are generalized as co-regularization style algorithms [22], where the disagreement between the functions of different views is taken as one part of the objective function to be minimized. Overall, by exploring the consistency and complementary properties of different views, multi-view learning is more effective than single-view learning.

One of the key challenges of multi-view classification comes from the fact that the incorporation of multiple views will bring much redundant and even conflicting information which is unfavorable for classifier learning. In order to tackle this problem, feature selection has been the focus of interest and much work has been done. Most of the existing studies can be categorized as filter models [17], [20] and embedded models based on sparsity regularization [7], [6], [26], [27]. While in this paper, we focus on wrapper models for feature selection. The problem of feature selection in the tensor product space is formulated as an integer quadratic programming problem in [21]. However, this method is limited to the interaction between two views and hard to extend to many views, since it directly selects features in the tensor product space resulting in the curse of dimensionality. [23] studies multi-view feature selection in the unsupervised setting.

We notice that support vector machine recursive feature elimination (SVM-RFE) can intelligently select discriminative features using the weight vector produced by support vector machine [9], but it can only be applied on a single-view data. In this paper, we use tensor product to organize multi-view features and study the problem of multi-view feature selection based on SVM-RFE and tensor techniques. Different from existing approaches, we leverage the correlations between the original data and the reconstructed tensors and develop a wrapper feature selection approach.

VI. Conclusion

In this paper, we studied the problem of multi-view feature selection. We explored tensor product to bring different views together in a joint space, and presented a dual method of tensor-based multi-view feature selection (DUAL-TMFS). Empirical studies in brain disease demonstrate the features selected by our proposed method yield better classification performance and are relevant to disease diagnosis.

Our proposed method has broad applicability for biomedical applications. Capabilities for simultaneous analysis of multiple feature subsets has transformative potential for yielding new insights concerning risk and protective relationships, for clarifying disease mechanisms, for aiding diagnostics and clinical monitoring, for biomarker discovery, for identification of new treatment targets and for evaluating effects of intervention.

VII. Acknowledgements

This work is supported in part by NSF through grants CNS-1115234, OISE-1129076, US Department of Army through grant W911NF-12-1-0066, NIH through grant R01-MH080636, and National Science Foundation of China (61472089).

References

[1] Elizabeth H Aylward, Paul D Brettschneider, Justin C McArthur, Gordon J Harris, Thomas E Schlaepfer, Jeffrey D Henderer, Patrick E Barta, Allen Y Tien, and Godfrey D Pearlson. Magnetic resonance imaging measurement of gray matter volume reductions in HIV dementia. *The American journal of psychiatry*, 152(7):987–994, 1995.

[2] Jason M Brenchley, Timothy W Schacker, Laura E Ruff, David A Price, Jodie H Taylor, Gregory J Beilman, Phuong L Nguyen, Alexander Khoruts, Matthew Larson, Ashley T Haase, et al. CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *The Journal of experimental medicine*, 200(6):749–759, 2004.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In *NIPS*, pages 396–404, 2009.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] Zheng Fang and Zhongfei Mark Zhang. Discriminative feature selection for multi-view cross-domain learning. In *CIKM*, pages 1321–1330. ACM, 2013.

[7] Yinfu Feng, Jun Xiao, Yueting Zhuang, and Xiaoming Liu. Adaptive unsupervised multi-view feature selection for visual concept recognition. In *ACCV*, pages 343–357, 2012.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv*, 2010.

[9] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[10] Lifang He, Xiangnan Kong, Philip S Yu, Ann B Ragin, Zhifeng Hao, and Xiaowei Yang. DuSK: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. In *SDM*. SIAM, 2014.

[11] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[12] Xiangnan Kong and Philip S Yu. Brain network analysis: a data mining perspective. *SIGKDD Explorations Newsletter*, 15(2):30–38, 2014.

[13] Xiangnan Kong, Philip S Yu, Xue Wang, and Ann B Ragin. Discriminative feature selection for uncertain graph classification. In *SDM*, 2013.

[14] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[15] Irma Mebane-Sims. 2009 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 2009.

[16] Michael W O'Dell, Deborah P Lubeck, Peter O'Driscoll, and Suzie Matsuno. Validity of the karnofsky performance status in an HIV-infected sample. *Journal of Acquired Immune Deficiency Syndromes*, 10(3):350–357, 1995.

[17] Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[18] RW Price, LG Epstein, JT Becker, P Cinque, Magnus Gisslén, L Pulliam, and JC McArthur. Biomarkers of HIV-1 CNS infection and injury. *Neurology*, 69(18):1781–1788, 2007.

[19] Ann B Ragin, Hongyan Du, Renee Ochs, Ying Wu, Christina L Sammet, Alfred Shoukry, and Leon G Epstein. Structural brain alterations can be detected early in HIV infection. *Neurology*, 79(24):2328–2334, 2012.

[20] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[21] Aaron Smalter, Jun Huan, and Gerald Lushington. Feature selection in the tensor product feature space. In *ICDM*, pages 1004–1009, 2009.

[22] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[23] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Unsupervised feature selection for multi-view data in social media. In *SDM*, 2013.

[24] Dacheng Tao, Xuelong Li, Xindong Wu, Weiming Hu, and Stephen J Maybank. Supervised tensor learning. *Knowledge and Information Systems*, 13(1):1–42, 2007.

[25] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *ICML*, pages 1065–1072, 2009.

[26] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, pages 352–360, 2013.

[27] Hua Wang, Feiping Nie, Heng Huang, and Chris Ding. Heterogeneous visual features fusion via sparse multimodal machine. In *CVPR*, pages 3097–3102, 2013.

[28] S Weis, H Haug, and H Budka. Neuronal damage in the cerebral cortex of AIDS brains: a morphometric study. *Acta neuropathologica*, 85(2):185–189, 1993.

[29] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv*, 2013.

[30] Jieping Ye, Kewei Chen, Teresa Wu, Jing Li, Zheng Zhao, Rinkal Patel, Min Bae, Ravi Janardan, Huan Liu, Gene Alexander, and Eric Reiman. Heterogeneous data fusion for Alzheimer's disease study. In *KDD*, pages 1025–1033. ACM, 2008.