Neural Process Lett (2010) 31:159–175
DOI 10.1007/s11063-010-9132-2

# A Novel Regularization Learning for Single-View Patterns: Multi-View Discriminative Regularization

**Zhe Wang · Songcan Chen · Hui Xue · Zhisong Pan**

Published online: 11 March 2010


**Abstract** The existing Multi-View Learning (MVL) is to discuss how to learn from patterns with multiple information sources and has been proven its superior generalization to the usual Single-View Learning (SVL). However, in most real-world cases there are just single source patterns available such that the existing MVL cannot work. The purpose of this paper is to develop a new multi-view regularization learning for single source patterns. Concretely, for the given single source patterns, we first map them into $M$ feature spaces by $M$ different empirical kernels, then associate each generated feature space with our previous proposed Discriminative Regularization (DR), and finally synthesize $M$ DRs into one single learning process so as to get a new Multi-view Discriminative Regularization (MVDR), where each DR can be taken as one view of the proposed MVDR. The proposed method achieves: (1) the complementarity for multiple views generated from single source patterns; (2) an analytic solution for classification; (3) a direct optimization formulation for multi-class problems without one-against-all or one-against-one strategies.

**Keywords** Discriminative Regularization · Multi-View Learning · Single source patterns · Multi-class problem · Classification

Z. Wang
Department of Computer Science and Engineering, East China University of Science and Technology, 200237 Shanghai, People's Republic of China

Z. Wang · S. Chen (✉) · H. Xue
Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, 210016 Nanjing, People's Republic of China
e-mail: s.chen@nuaa.edu.cn

H. Xue
School of Computer Science and Engineering, Southeast University, 210096 Nanjing, People's Republic of China

Z. Pan
Institute of Command Automation, PLA University of Science and Technology, 210007 Nanjing, People's Republic of China

## 1 Introduction

Since the pattern is the dealt object of the classifier, it is important to consider the prior knowledge of patterns in designing classifiers [13]. In practice, patterns can be obtained from single or multiple information sources. If each information source is taken as one view, accordingly there are two kinds of patterns, i.e. single-view patterns and multi-view patterns. Each information source may induce one attribute set for patterns. Thus, single-view patterns are composed of single attribute set and multi-view patterns are composed of multiple attribute sets. Correspondingly, the learning on single-view and multi-view patterns can be sorted into Single-View and Multi-View Learning (SVL and MVL), respectively. In the literature [5,30], it has been demonstrated that co-training (one typical MVL approach) has a superior generalization ability to its corresponding SVL for semi-supervised learning. Given patterns that are composed of two naturally-split attribute sets (two views), co-training requires the assumption that two views given the class are conditionally independent. Here, the independence assumption is guaranteed by the patterns composed of two naturally-split attribute sets.

Regularization learning [7,8,10,17,39] is viewed as one effective method for improving the generalization performance of classifiers. It has a rich history which can date back to the theory of ill-posed problem [27,39,40]. By incorporating the right amount of prior information into the formulation, regularization techniques are shown to be powerful in making the solution stable [7,19]. Regularization theory is introduced to the machine learning community on the premise that the learning can be viewed as a multivariate functional fitting problem, and also is successfully applied to the classifier learning [7,32].

The goal of this paper is: (1) to develop a new supervised MVL for single-view patterns; (2) to incorporate the proposed MVL in regularization learning for a superior classification performance, whose underlying motivations and contributions are as follows:

- The proposed MVL can deal with single-view patterns without the independence assumption. In most real-world applications, it is not well satisfied for the independence assumption of the attribute sets since there are only single-view patterns available. In that case, the existing MVL can not effectively work [2,49,50]. However, it is this fact that motivates us to develop a new MVL on single-view patterns.
- The proposed MVL adopts multiple kernels. It is well-known that the types and the parameters of the kernels must be selected in practice. For a given application, there may be multiple kernels as the candidates which can possess different types and parameters. The kernel selected from the candidates can yield a model with good performance. Such a selection, equivalently to model selection, can usually be achieved by some methods of optimizing kernels such as Cross Validation (CV) or Leave-One-Out (LOO) [6,26]. However, these methods are computationally expensive when dealing with a large number of kernel types or parameters. Even the kernel selected by these optimization methods also can not be guaranteed optimality in some cases. Further, since the selected kernel is single and fixed, it can only characterize the geometrical structure of some aspects for the input data and, thus, is not always a good fit for the applications which involve multiple, heterogeneous data sources, which is validated in the literature [37]. To this end, a method based Multiple Kernel Learning named MKL was proposed [4,11,16,20,21,31,44]. They showed the necessity to consider multiple kernels or the combination of kernels rather than a single fixed kernel. Generally, MKL tries to form an ensemble of kernels so as to yield a good fit for a certain application. It has been proven that MKL can offer some needed flexibility and well manipulate the case that involves multiple, heterogeneous data sources [1,3,37]. Since MKL considers multiple kernels, it can be effectively employed

for the heterogeneous data sources under the common framework of kernel learning. To a certain extent, MKL also relaxes the model selection about kernels. Thus, we adopt multiple kernels in the multiple view learning framework here.

- The proposed MVL first adopts multiple empirical kernel mappings [35,45] for the given single-view patterns. Then it synthesizes different mappings so as to achieve the complementarity among the generated views and get a superior classification performance to the original SVL, where each associated empirical kernel mapping is taken as one view of the original single-view patterns. Each view is expected to be able to exhibit some geometrical structure of the original patterns from its own perspective such that all the views can complement each other. In practice, the complementarity among multiple views is achieved by the following so-called Inter-Function Similarity Loss term $R_{IFSL}$ [44]:

$$R_{IFSL}(x) = \sum_{l=1}^{M} \left( f_l(x) - \sum_{j=1}^{M} \alpha_j f_j(x) \right)^2, \tag{1}$$

  where $x \in \mathbb{R}^n$ is a given single-view pattern, $f_j$ is a classifier learnt from the $j$th kernel mapping space of the original patterns, and $\alpha_j \geq 0$, $\sum_{j=1}^{M} \alpha_j = 1$, $\alpha_j$ denotes the importance of the corresponding view. It can be found that for a given pattern, $R_{IFSL}$ expects to make all the $M$ classifiers $f_l$ achieve as much agreement on their outputs as possible.
- The proposed MVL adopts our previous work [47] of Discriminative Regularization (DR) as $f_j$ in the term $R_{IFSL}$, and thus is named as Multi-view Discriminative Regularization (MVDR). MVDR inherits the advantages of DR and owns: (1) an analytic solution for classification; (2) a direct optimization formulation for multi-class problems without one-against-all or one-against-one strategies. Meanwhile, since the proposed MVDR considers multiple views generated from the original pattern and achieves the complementarity among these views, it has a superior classification performance to the original DR, which is validated in the experiments of this paper.
- The proposed MVL is applied into *supervised* problems and experimentally shows that a weaker correlation between the views of the proposed method leads to a performance improvement. Most of the existing MVL works along *semi-supervised* problems [5,28,30]. But this paper changes it and applies the MVL technique into supervised problems. Meanwhile, the literature [43] has theoretically and experimentally given that if the base learners of co-training style algorithms have enough differences in semi-supervised cases, an improved performance can be got. This paper extends the similar conclusion of the literature [43] to supervised cases and experimentally gives that a weaker correlation between the views can lead to a superior performance.

This paper is organized as follows. Section 2 describes the related work in MVL. Section 3 reviews our previous work DR. The architecture of the proposed MVDR is given in Sect. 4. Section 5 reports the experimental results on some benchmark data sets and shows the feasibility and effectiveness of the proposed MVDR. Finally, the conclusion is given.

## 2 Related Work

One typical example of the existing MVL is web-page classification [5], where each web page can be represented by either the words on itself (view one) or the words contained in anchor texts of inbound hyperlinks (view two). Blum and Mitchell [5] design a co-training algorithm on the labeled and unlabeled web pattern sets composed of the two naturally-split views. For

the co-training style algorithm, two classifiers are incrementally built with the corresponding views on the labeled web set. On each cycle, each classifier labels the unlabeled webs and picks those with the highest confidence into the labeled set. The co-training process repeats until the terminated condition is satisfied. It is well-known that the co-training algorithm requires two assumptions: (1) the compatibility assumption that the base classifiers in each view farthest agree on labels of web patterns and (2) the independence assumption that the different views given the class are conditionally independent. But in most cases, it is hard to satisfy the independence assumption due to the nonexistence of naturally-split attribute sets (naturally-split views) such as the single-view patterns. Thus Nigam and Ghani [30] experimentally explore the co-training algorithm with or without the independence assumption. They demonstrate that the co-training algorithm with a natural split of the attributes outperforms the one without, and further propose a semi-supervised, multi-view algorithm co-EM that is a probabilistic version of co-training and outperforms co-training. Moreover, Muslea et al. [28] incorporate active learning in co-EM, and present an algorithm named co-EMT that outperforms both co-training and co-EM and has a robustness in view-correlation cases to some extent.

Although both co-EMT and co-EM have the superior generalization to co-training, all these algorithms can not effectively work on the patterns with the non-naturally split attributes, especially the single-view patterns. In order to solve the problem, Zhang et al. [49] design an algorithm called Correlation and Compatibility based Feature Partitioner (CCFP) to automate multi-view detection, where the attributes of patterns can be partitioned into two views that are low correlated, compatible and sufficient enough. But, as the authors themselves said in [49], CCFP has two limitations: (1) the two views must have the same number of attributes and certain correlation; (2) it is hard to get the optimal parameters of CCFP. Farquhar et al. [15] present a process named SVM-2K that combines Kernel Canonical Correlation Analysis (KCCA) [18] by Support Vector Machine (SVM) [42] on two views. SVM-2K utilizes the multi-kernel trick on the single-view patterns, where for the same pattern the two views are generated through two feature projections $\phi_A$ and $\phi_B$ with their corresponding kernels $k_A$ and $k_B$. However, due to SVM itself, SVM-2K also suffers from similar problems such as the scalability to the number of the patterns and time-consuming Quadratic Programming (QP). On the other hand, rather than dealing with the single-view patterns themselves, the democratic co-learning [50] runs different algorithms on the single-view patterns, whose motivations are that different learning algorithms yield different inductive biases and that better performance can be made by the voted majority. However, in the democratic co-learning, how to select those base learning algorithms is still a problem due to lack of a measurable selection criterion.

Compared with CCFP, SVM-2K and the democratic co-learning, the proposed MVDR has the following advantages: (1) it does not need to split the attributes of the original single-view patterns but just maps the original single-view patterns into $M$ feature spaces with $M$ empirical kernels, respectively; (2) it can achieve the complementarity among the so-generated feature spaces through introducing the term $R_{IFSL}$; (3) it employs our previous work of DR as the base learner in the individual feature spaces, and thus owns a nice analytic solution and a direct optimization formulation for multi-class problems.

## 3 Discriminative Regularization

It has be demonstrated that the traditional regularization learning usually just considers one side of classification problems. Regularization Network (RN) [19] only emphasizes the smoothness of the classifier, and does not sufficiently incorporate the prior intra-class

and inter-class information into its formulation which is vital for classification. Generalized Radial Basis Function Network (GRBFN) [32], as an approximation to RN, actually just incorporates the intra-class information generated from the clusters into the traditional regularization learning. But, GRBFN still partially neglects the inter-class information which is crucial for classification. SVM uses the hinge-loss function and thus emphasizes the prior inter-class discriminative knowledge more than GRBFN. Furthermore, Regularized Least Squares (RLS) method [33] is established by minimizing a regularized function directly in a Reproducing Kernel Hilbert Space (RKHS). RLS is proved to have a similar performance to SVM [48]. However, both RLS and SVM do not take the intra-class information into account yet and thus do not sufficiently use the prior data structural knowledge, which may influence classification effectiveness to some degree. Discriminative Regularization (DR) [47] was proposed to improve the traditional regularization for classification, but does not change the original formulation. DR directly introduces the prior not only intra-class but also inter-class information into the objective function as discriminative knowledge [47].

Suppose that we are given the binary-class problem $\{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^n \times \{-1, +1\}$, where $y_i$ is the class label of the training pattern $x_i$. The linear discriminant function of DR is given as follows

$$f(x) = w^T x + b, \tag{2}$$

where $w \in \mathbb{R}^n$ is the weight vector and $b \in \mathbb{R}$ is a bias. $w$ and $b$ is optimized by the following objective function

$$\min_{w,b} \frac{1}{2} \sum_{i=1}^{N} \left[ y_i - \left( w^T x_i + b \right) \right]^2 + \frac{1}{2} w^T \left[ \eta S_w^e + (\eta - 1) S_b^e \right] w, \tag{3}$$

where

$$S_w^e = \sum_{k=1}^{2} \frac{1}{N_k} \sum_{i=1}^{N_k} \left( x_i^{(k)} - \overline{x}^{(k)} \right) \left( x_i^{(k)} - \overline{x}^{(k)} \right)^T$$

$$S_b^e = \sum_{k=1}^{2} \sum_{p \neq k} \left( \overline{x}^{(k)} - \overline{x}^{(p)} \right) \left( \overline{x}^{(k)} - \overline{x}^{(p)} \right)^T,$$

$N_k$ is the number of the $k$th class patterns, $x_i^{(k)}$ denotes the $i$th pattern of the $k$th class, $\overline{x}^k$ denotes the average pattern of the $k$th class, and $\eta$ is the parameter that regulates the relative significance of the intra-class compactness versus the inter-class separability, $0 \leq \eta \leq 1$. The second term of the formulation (3) is exactly called as Discriminative Regularization Term that contains both the prior intra-class and inter-class information.

It should be stated that both $S_w^e$ and $S_b^e$ are much similar to the well-known "within-class scatter matrix" and "between-class scatter matrix" in Linear Discriminant Analysis (LDA), respectively [24]. Hence actually, the regularization term in DR is naturally coincident with the formulation of Maximum Margin Criterion (MMC) [23]. Although DR is a classifier learning method rather than traditional dimensionality reduction, i.e., the optimized $w$ is actually the weight vector in the classifier functional rather than the projection vector, DR more likely provides us a brand-new viewpoint of combining regularization with supervised dimensionality reduction methods effectively. The general goal of supervised dimensionality reduction methods, such as LDA and MMC, is to find an orientation in which the projected samples are well separated [12], which is much similar to the intuitive motivation in DR. Hence through introducing these methods into the regularization framework as a

regularization term, DR virtually provides a general way to incorporate the prior information into the formulation of designing a new classifier, which extends the traditional regularization to classification. The detailed description about DR can be found in [47].

## 4 Multiple Views of Discriminative Regularization

In the proposed MVL, given the single-view training patterns $\{(x_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}^n \times \{C_1, \ldots, C_c\}$, we can map each pattern $x_i$ from the input space $\mathcal{X}$ into $M$ feature spaces $\{\mathcal{F}_l^{n_l}\}_{l=1}^M$ with $M$ kernels, i.e., $\Phi_l : \mathcal{X} \to \mathcal{F}_l^{n_l}, l = 1, \ldots, M$. Each generated feature space $F_l^{n_l}$ has $n_l$ dimension. The aim of the proposed MVL is to use all the $M$ generated feature spaces and achieve the complementarity among all the feature spaces.

In the literature [35,36], the mapping $\Phi$ also called the Implicit Kernel Mapping (IKM) is *implicitly* represented by specifying a kernel function as the inner product between each pair of samples in the feature space. For the sample set $\{x_i\}_{i=1}^N$, $X$ denotes the $N \times n$ sample matrix where each row is the vector $x_i^T$. $K = [ker_{ij}]_{N \times N}$ denotes the $N \times N$ kernel matrix where $ker_{ij} = \Phi(x_i) \cdot \Phi(x_j) = ker(x_i, x_j)$. $K$ is a symmetrical positive-semidefinite matrix. Conversely, the mapping $\Phi$ in this paper, is given in an *explicit* form as describe in [35,45]. If the rank of $K$ is $r$, the kernel matrix $K$ can be decomposed as

$$K_{N \times N} = Q_{N \times r} \Lambda_{r \times r} Q_{r \times N}^T, \tag{4}$$

where $\Lambda$ is a diagonal matrix consisting of the $r$ positive eigenvalues of $K$, and $Q$ consists of the corresponding orthonormal eigenvectors. Then, the explicit mapping also called the Empirical Kernel Mapping (EKM) in this paper, is given as

$$\Phi^e : \mathcal{X} \to \mathcal{F}^r$$

$$x \to \Lambda^{-1/2} Q^T [ker(x, x_1), \ldots, ker(x, x_N)]^T. \tag{5}$$

Let $B = KQ\Lambda^{-1/2}$, and then the dot product matrix of $\{\Phi^e(x_i)\}_{i=1}^N$ generated by EKM can be calculated as

$$BB^T = KQ\Lambda^{-1/2}\Lambda^{-1/2}Q^T K = K. \tag{6}$$

Equation 6 of EKM is exactly equal to the kernel matrix (4) of IKM. Thus the mapped samples, respectively, generated by EKM and IKM have the same geometrical structure. In [35,45], it is shown that comparing EKM with IKM, the former is easier to access and easier to study the adaptability of a kernel to the input space than the latter. That is why we select EKM here.

This paper generates $M$ different feature spaces with $M$ EKMs, where each feature space is taken as one view of the given training patterns. Each view only shows one-facet structural information of the original patterns. Thus, the learning in one certain feature space might be just local or partial. The proposed MVL is expected to employ all the generated feature spaces and complement all the individual learnings in $M$ feature spaces. Such a complementarity in the proposed MVL can be achieved through utilizing the prior knowledge in the training patterns, which is also validated in the literature [22]. It can be found that though $x_i$ can be mapped into different feature patterns $\left\{\Phi_l^e(x_i)\right\}_{l=1}^M$, $\left\{\Phi_l^e(x_i)\right\}_{l=1}^M$ still share a common class label $y_i$. Therefore, denote $f_l$ as the classifier learnt from the $l$th feature space $\mathcal{F}_l$, and then

the outputs of all the classifiers $\{f_l\}_{l=1}^{M}$ on $x_i$ should achieve as much agreement as possible, which is here characterized by the Inter-Function Similarity Loss term

$$R_{IFSL} = \sum_{l=1}^{M} \left[ f_l(x_i) - \sum_{j=1}^{M} \alpha_j f_j(x_i) \right]^2$$
$$\alpha_j \geq 0, \quad \sum_{j=1}^{M} \alpha_j = 1.$$

DR is used to construct the classifier $f_l$ in each view $\mathcal{F}_l$. Further, we will give the formulation of the proposed MVL called MVDR in the next section.

### 4.1 Binary-Class Problem

This section gives the formulation of the proposed MVDR for binary-class problem. The original single-view patterns $\{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^n \times \{-1, +1\}$ are mapped into $\left\{ \left\{ \Phi_l^e(x_i) \right\}_{l=1}^{M} \right\}_{i=1}^{N}$ with $M$ empirical kernels as shown in (5). The classifier $f_l$ of each view $\Phi_l^e$ in the proposed MVDR has the linear formulation

$$f_l(x) = w_l^T \Phi_l^e(x) + b_l \tag{7}$$

as in DR. Then, the decision function of MVDR is formed as

$$F(x) = \sum_{l=1}^{M} \alpha_l \left[ w_l^T \Phi_l^e(x) + b_l \right], \tag{8}$$

where $\alpha_l \geq 0$, $\sum_{l=1}^{M} \alpha_l = 1$.

As a result, the optimization problem of MVDR is characterized as below

$$\min_{w_l, b_l} J = R_{emp} + R_{DR} + \lambda R_{IFSL}, \tag{9}$$

where $R_{emp}$, $R_{DR}$ are the empirical risk term and the discriminant term of $M$ views, respectively, and $R_{IFSL}$ is the inter-function similarity loss term. $R_{emp}$, $R_{DR}$, and $R_{IFSL}$ are, respectively, defined as

$$R_{emp} = \frac{1}{2} \sum_{l=1}^{M} \sum_{i=1}^{N} \left[ y_i - \left( w_l^T \Phi_l^e(x_i) + b_l \right) \right]^2, \tag{10}$$

$$R_{DR} = \frac{1}{2} \left[ \eta \sum_{l=1}^{M} w_l^T S_w^l w_l + (\eta - 1) \sum_{l=1}^{M} w_l^T S_b^l w_l \right], \tag{11}$$

$$R_{IFSL} = \frac{1}{2} \sum_{i=1}^{N} \sum_{l=1}^{M} \left\{ \left[ w_l^T \Phi_l^e(x_i) + b_l \right] - \sum_{j=1}^{M} \alpha_j \left[ w_j^T \Phi_j^e(x_i) + b_j \right] \right\}^2, \tag{12}$$

where

$$S_w^l = \sum_{k=1}^{2} \frac{1}{N_k} \sum_{i=1}^{N_k} \left(\Phi_l^e\left(x_i^{(k)}\right) - \Phi_l^e\left(\overline{x}^{(k)}\right)\right)\left(\Phi_l^e\left(x_i^{(k)}\right) - \Phi_l^e\left(\overline{x}^{(k)}\right)\right)^T,$$

$$S_b^l = \sum_{k=1}^{2} \sum_{p\neq k} \left(\Phi_l^e\left(\overline{x}^{(k)}\right) - \Phi_p^e\left(\overline{x}^{(p)}\right)\right)\left(\Phi_l^e\left(\overline{x}^{(k)}\right) - \Phi_l^e\left(\overline{x}^{(p)}\right)\right)^T,$$

$\Phi_l^e(\cdot), w_l \in \mathbb{R}^{n_l}, b_l \in \mathbb{R}$. Both $R_{emp}$ and $R_{DR}$ characterize the $M$ DR learnings in their corresponding feature spaces. $R_{IFSL}$ achieves the complementarity among the $M$ DRs. For conveniently processing $w_l$ and $b_l$, we reformulate $R_{emp}$, $R_{DR}$, and $R_{IFSL}$ in matrix form:

$$R_{emp} = \frac{1}{2}\left(Y - \mathbf{X}^T u\right)^T \left(Y - \mathbf{X}^T u\right) + \frac{M-1}{2} Y^T Y, \tag{13}$$

$$R_{DR} = \frac{1}{2} u^T \left[\eta S_w^e + (\eta - 1) S_b^e\right] u, \tag{14}$$

$$\begin{aligned} R_{IFSL} &= \frac{1}{2} \sum_{l=1}^{M} \left(u_l^T X_l - u^T \Lambda \mathbf{X}\right)\left(u_l^T X_l - u^T \Lambda \mathbf{X}\right)^T \\ &= \frac{1}{2}\left(u^T \mathbf{X}\mathbf{X}^T u + M u^T \Lambda \mathbf{X}\mathbf{X}^T \Lambda u - 2u^T \mathbf{X}\mathbf{X}^T \Lambda u\right), \end{aligned} \tag{15}$$

where

$$\begin{aligned} Y &= [y_1, \ldots, y_n]^T, \\ u_l &= \left[w_l^T, b_l\right]^T, \\ u &= \left[u_1^T, \ldots, u_M^T\right]^T, \end{aligned}$$

$\Lambda$ is a diagonal matrix with its diagonal elements in the sequence being

$$\begin{aligned} &\alpha_1^1 \ldots \alpha_1^{n_1+1}, \ldots, \alpha_l^1 \ldots \alpha_l^{n_l+1}, \ldots, \alpha_M^1 \ldots \alpha_M^{n_M+1}, \\ &X_l = \begin{bmatrix} \Phi_l^e(x_1) & \ldots & \Phi_l^e(x_N) \\ 1 & \ldots & 1 \end{bmatrix}, \\ &\mathbf{X} = [X_1; \ldots; X_M]. \end{aligned}$$

Denote $\mathbf{X} = [z_1, \ldots, z_N]$, then

$$S_w^e = \sum_{k=1}^{2} \frac{1}{N_k} \sum_{i=1}^{N_k} \left(z_i^{(k)} - \overline{z}^{(k)}\right)\left(z_i^{(k)} - \overline{z}^{(k)}\right)^T,$$

$$S_b^e = \sum_{k=1}^{2} \sum_{p\neq k} \left(\overline{z}^{(k)} - \overline{z}^{(p)}\right)\left(\overline{z}^{(k)} - \overline{z}^{(p)}\right)^T.$$

Thus, to get the minimizer of the objective function $J$ in the Eq. 9, we make the gradient of $J$ with respect to $u = \left[u_1^T, \ldots, u_M^T\right]^T \left(u_l = \left[w_l^T, b_l\right]^T\right)$ be zero and get

$$\frac{\partial J}{\partial u} = \frac{\partial R_{emp}}{\partial u} + \frac{\partial R_{DR}}{\partial u} + \lambda \frac{\partial R_{IFSL}}{\partial u} = 0. \tag{16}$$

Then, the Eq. 17 can be induced through settling the Eq. 16 as following

$$\left\{(1+\lambda)A + \left[\eta S_w^e + (\eta - 1)S_b^e\right] + \lambda M \Lambda A \Lambda - \lambda \left(A\Lambda + \Lambda A\right)\right\} u = \mathbf{X}Y, \tag{17}$$

where $A = \mathbf{X}\mathbf{X}^T$. An analytic solution to the $u$ can be obtained.

### 4.2 Multi-Class Problem

In the $c$-class problem ($c \geq 2$), we adopt the vector-labeled outputs that can make the computational complexity independent of the number of classes and require no more computation than a single binary classifier [14]. Furthermore, Szedmak and Shawe-Taylor [38] presents that this technique of the vector-labeled outputs does not diminish classification performance but in some cases can improve it, relatively to one-against-one and one-against-all for multi-class problems. Therefore, this paper codes the class labels with the one-of-$c$ rule. If $x_i$ belongs to the $k$th class, then its label $y_i = [0 \ldots 1 \ldots 0]^T \in \mathbb{R}^c$, where the $k$th element is 1 and the other elements are 0. Then the classifier (8) of the proposed MVDR for the $c$-class problem can be formulated as

$$F(x) = \sum_{l=1}^{M} \alpha_l \left[\mathbf{W}_l^T \Phi_l^e(x) + \mathbf{b}_l\right], \tag{18}$$

where $\mathbf{W}_l \in \mathbb{R}^{n_l \times c}$, $\mathbf{b}_l \in \mathbb{R}^c$. Correspondingly, the objective function of the proposed MVDR for the $c$-class problem is formulated as

$$\min_{\mathbf{W}_l, \mathbf{b}_l} J = R_{emp} + R_{DR} + \lambda R_{IFSL}, \tag{19}$$

where

$$R_{emp} = \frac{1}{2} tr\left[\left(\mathbf{Y} - \mathbf{U}^T\mathbf{X}\right)^T \left(\mathbf{Y} - \mathbf{U}^T\mathbf{X}\right)\right], \tag{20}$$

$$R_{DR} = \frac{1}{2}\left[\eta \widetilde{S}_w^e + (\eta - 1)\widetilde{S}_b^e\right], \tag{21}$$

$$\begin{aligned} R_{IFSL} &= \frac{1}{2}\sum_{l=1}^{M} tr\left[\left(U_l^T X_l - U^T \Lambda \mathbf{X}\right)\left(U_l^T X_l - U^T \Lambda \mathbf{X}\right)^T\right] \\ &= \frac{1}{2}\left(\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} + M\mathbf{U}^T\Lambda\mathbf{X}\mathbf{X}^T\Lambda\mathbf{U} - 2\mathbf{U}^T\mathbf{X}\mathbf{X}^T\Lambda\mathbf{U}\right), \end{aligned} \tag{22}$$

$tr[\cdot]$ is a matrix trace operation. In this case,

$$\mathbf{Y} = [y_1, \ldots, y_N] \in \mathbb{R}^{c \times N}, \quad y_i \in \mathbb{R}^c,$$
$$\mathbf{U} = \left[U_1^T, \ldots, U_M^T\right]^T, \quad U_l = \left[\mathbf{W}_l^T, \mathbf{b}_l\right]^T,$$

**Table 1** Algorithm MVDR

**Input**: $\{x_i, y_i\}_{i=1}^N$; the $M$ candidate kernels $\left\{ker_l\left(x_i, x_j\right)\right\}_{l=1}^M$
**Output**: the solution in the binary problem: $w_l, b_l$ (the multi-class problem: $\mathbf{W}_l, \mathbf{b}_l$), $l = 1 \ldots M$
1. Explicitly map $\{x_i\}_{i=1}^N$ into $\left\{\Phi_1^e(x_i), \ldots, \Phi_l^e(x_i), \ldots, \Phi_M^e(x_i)\right\}_{i=1}^N$
by $M$ kernels as shown in (5);
2. Set $u = \left[u_1^T, \ldots, u_M^T\right]^T$, $u_l = \left[w_l^T, b_l\right]^T$ (the multi-class problem :
$\mathbf{U} = \left[U_1^T, \ldots, U_M^T\right]^T$, $U_l = \left[\mathbf{W}_l^T, \mathbf{b}_l\right]^T$),
then $u$ ($\mathbf{U}$) can be got through (17) (the multi-class problem: (24)).

both $\mathbf{X}$ and $\Lambda$ follow the definition of the binary-class problem. Denote $\mathbf{X} = [z_1, \ldots, z_N]$ again, then

$$\widetilde{S}_w^e = \sum_{k=1}^{c} \frac{1}{N_k} \sum_{i=1}^{N_k} \left(z_i^{(k)} - \overline{z}^{(k)}\right)^T \mathbf{U}\mathbf{U}^T \left(z_i^{(k)} - \overline{z}^{(k)}\right),$$
$$\widetilde{S}_b^e = \sum_{k=1}^{c} \sum_{p \neq k} \left(\overline{z}^{(k)} - \overline{z}^{(p)}\right)^T \mathbf{U}\mathbf{U}^T \left(\overline{z}^{(k)} - \overline{z}^{(p)}\right).$$

Similarly, to get the minimizer of the objective function $J$ in the multi-class problem (19), we zero the gradient of $J$ of (19) with respect to $\mathbf{U} = \left[U_1^T, \ldots, U_M^T\right]^T$ $\left(U_l = \left[\mathbf{W}_l^T, \mathbf{b}_l\right]^T\right)$ and get

$$\frac{\partial J}{\partial \mathbf{U}} = \frac{\partial R_{emp}}{\partial \mathbf{U}} + \frac{\partial R_{DR}}{\partial \mathbf{U}} + \lambda \frac{\partial R_{IFSL}}{\partial \mathbf{U}} = 0. \tag{23}$$

Then, the Eq. 24 can be induced through settling the Eq. 23 as following

$$\left\{(1+\lambda)A + \left[\eta S_w^e + (\eta - 1)S_b^e\right] + \lambda M \Lambda A \Lambda - \lambda(A\Lambda + \Lambda A)\right\}\mathbf{U} = \mathbf{X}Y^T, \tag{24}$$

where

$$A = \mathbf{X}\mathbf{X}^T,$$
$$S_w^e = \sum_{k=1}^{c} \frac{1}{N_k} \sum_{i=1}^{N_k} \left(z_i^{(k)} - \overline{z}^{(k)}\right)\left(z_i^{(k)} - \overline{z}^{(k)}\right)^T,$$
$$S_b^e = \sum_{k=1}^{c} \sum_{p \neq k} \left(\overline{z}^{(k)} - \overline{z}^{(p)}\right)\left(\overline{z}^{(k)} - \overline{z}^{(p)}\right)^T.$$

Thus, we can obtain an analytic solution to the weight matrix for classifier of the proposed MVDR in the multi-class problem.

Table 1 lists the procedure of the proposed MVDR in both binary and multi-class problems. From this table, it can be found that the proposed MVDR has two advantages: (1) an analytic solution to the optimization problem; (2) a direct optimization formulation for multi-class problems without one-against-all or one-against-one strategies.

## 5 Experiments

The used single-view patterns in our experiments are the synthetic data and UCI data sets [29], respectively. The used candidate kernels for all the implemented algorithms are: linear kernel $ker(x_i, x_j) = x_i^T x_j$; RBF kernel $ker(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$ where $\sigma = \nu\overline{\sigma}$, $\nu$ is selected from $\{0.01, 0.1, 1, 10, 100\}$, $\overline{\sigma}$ is set to the average value of all the $l_2$-norm distances $\|x_i - x_j\|_2, i, j = 1 \ldots N$ as used in [41]; and polynomial kernel $ker(x_i, x_j) = \left(x_i^T x_j + 1\right)^d$ where $d$ is selected from $\{2, 3, 4, 5\}$, respectively. Without any prior knowledge, the parameter $\alpha_l, l = 1 \ldots M$ of the proposed MVDR is set to $\frac{1}{M}$, i.e., each view owns the same importance. The range of the parameter $\eta$ for DR is $\{0.001, 0.01, 0.1, 0.5, 0.7, 0.99\}$. The parameter $\lambda$ for MVDR is from $10^{-3}$ to $10^2$ with each step by multiplying 10. The classification performances of all the algorithms here are reported by Monte Carlo cross validation (MCCV) [46] that randomly splits the pattern set into two parts (the training and testing sets), and then repeats the procedure $T$ times. Here, $T$ is set to 10.

### 5.1 Synthetic Data

Figure 1 demonstrates the complementarity of the proposed MVDR on the synthetic data sets, where the data in two classes ('○' vs. '+') appear as two banana shaped distributions. The data are uniformly distributed along the bananas and are superimposed with a normal distribution with standard deviation in all directions. Figure 1a–c give the boundaries of DR with linear, polynomial, and RBF kernels in the synthetic data, respectively. In contrast, Fig. 1d gives the boundary of the proposed MVDR with the same linear, polynomial, and RBF kernels as those used in Fig. 1a–c. Furthermore, the training and testing accuracies are labeled in the right-bottom corners in their corresponding sub-figures.

From this figure, it can be found that: (1) the proposed MVDR has a more accurate decision boundary that well sketches the real contour of the '+' patterns; (2) DR with the linear kernel clearly gives an under-fitting decision boundary that only gives a general trend of the data distribution; (3) DR with the polynomial or RBF kernels has a better classification performance than DR with the linear kernel respectively, but still fails in classifying some certain patterns that lie in the boundary area; (4) the proposed MVDR employs multiple kernels and exhibits the best classification accuracy.

Further, Fig. 1a–c showed the decision boundaries for linear, polynomial and RBF kernels while Fig. 1d showed the decision boundary combining the above three. Some '+' samples are to the left of the decision boundary for all linear, polynomial and RBF kernels in Fig. 1a–c. That is to say, none of the three kernels can learn these '+' samples well. However, these samples were to the right of the boundary in Fig. 1d where the three kernels were combined. To analyze the reason, it should be stated that the classifier functions of DR with linear, polynomial and RBF kernels in Fig. 1a–c are different from those of MVDR with the combination of linear, polynomial and RBF kernels in Fig. 1d due to the difference between the solutions of DR and MVDR. As stated in Sect. 4, the proposed MVDR is not simply combined by the separate DR. The $\mathbf{W}_l^T, \mathbf{b}_l, l = 1 \ldots M$ in the MVDR are optimized in one learning processing and play an influence for each other. Therefore, although none of the three kernels in DR can learn these '+' samples well, these '+' samples can also be learned right by MVDR in Fig. 1. That is to say that the three sub-classifiers in MVDR are different from that the three classifiers of DR. It is thus not contradictory to the assumption that these kernels in MVDR are complementary. To further validate the proposed MVDR, we

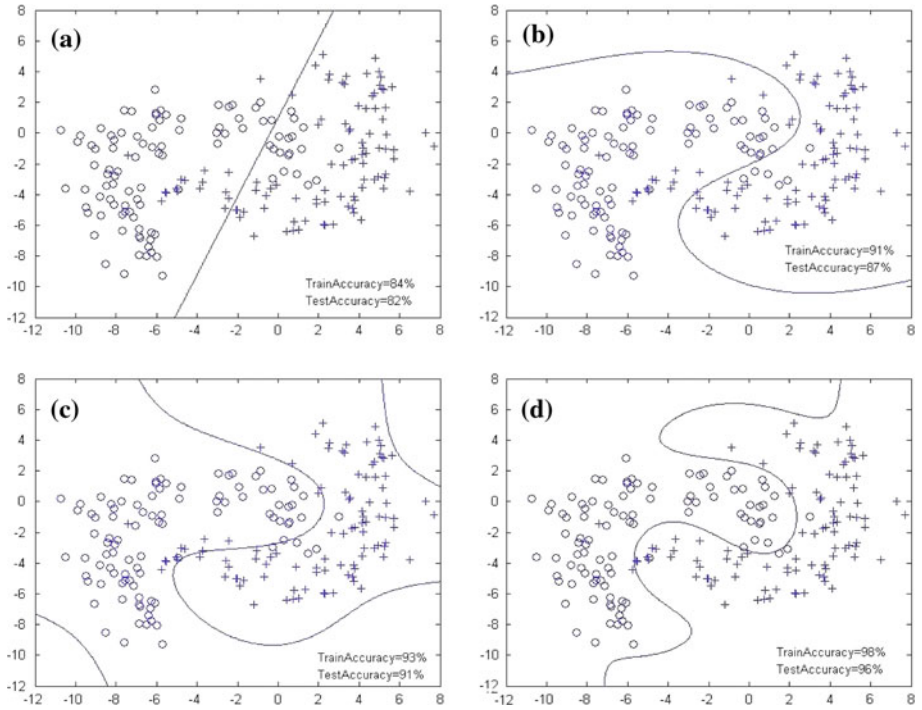


**Fig. 1** The discriminant boundaries in the two-banana data set: **a** DR with linear kernel; **b** DR with polynomial kernel; **c** DR with RBF kernel; **d** MVDR with the same linear, polynomial, and RBF kernels as (**a**)–(**c**)

will compare it with DRMV that separately carries out the $M$ DR algorithms in the $M$ feature spaces, respectively, and then combines their outputs by the majority voting technique in the next section.

## 5.2 UCI Data Sets

### *5.2.1 Classification Performance*

This section implements the proposed MVDR on UCI data sets to further validate its effectiveness. Simultaneously, this section also carries out the DR algorithm based on the single kernel and two kinds of combinations (denoted as DRMV and DRFE, respectively). The DRMV separately carries out the $M$ DR algorithms in the $M$ feature spaces, respectively, and then combines their outputs by the majority voting technique. The DRFE first concatenates the $M$ transformed feature vectors into one single ensemble vector, and then implements the DR algorithm with the ensemble vector. In addition, the multiple kernel learning algorithm denoted as MKL [34] is also compared with the proposed method. All the implemented algorithms MVDR, DRFE, DRMV and MKL [34] adopt the same empirical kernels where $M$ is set to 3 or 5 on the used data sets. The results of the algorithm DR are given in the optimal kernel case through MCCV. We first give the experimental results of the DR with different kernels (views) and SVM with RBF kernels. We list the results in Table 2. From this table, we can find that the proposed MVDR has a significant superiority to the

**Table 2** Classification accuracy comparison between the algorithms MVDR, DR, and SVM

| Data sets | DR | | | MVDR | SVM |
|---|---|---|---|---|---|
| | Linear | poly | RBF | combination | RBF |
| Sonar | 0.7231 | 0.6481 | 0.7296 | **0.7639** | 0.7333 |
| Iono. | 0.6393 | 0.6707 | 0.8033 | 0.9047 | **0.9426** |
| Hous. | 0.7511 | 0.7819 | 0.7511 | **0.9267** | 0.9239 |
| Echo. | 0.6045 | 0.6239 | 0.6134 | 0.6298 | **0.8776** |
| Shut. | 0.5714 | 0.6285 | 0.6142 | **0.6714** | 0.5714 |
| Glas. | 0.7295 | 0.6514 | 0.7733 | 0.8695 | **0.8761** |
| Soy. | 0.9956 | 0.9956 | **1** | **1** | 0.9173 |
| Der. | 0.2888 | 0.4361 | 0.2988 | 0.4716 | **0.4733** |
| Lens. | 0.2923 | 0.3384 | 0.3461 | 0.3769 | **0.5846** |
| Cmc | 0.4088 | 0.4517 | 0.4774 | 0.5064 | **0.5168** |
| Wine | 0.3103 | 0.6896 | 0.5745 | **0.9056** | 0.8443 |
| Lung. | 0.4733 | 0.4 | 0.48 | **0.5066** | 0.4133 |

Bold values indicate the best accuracy of the algorithms on each data set

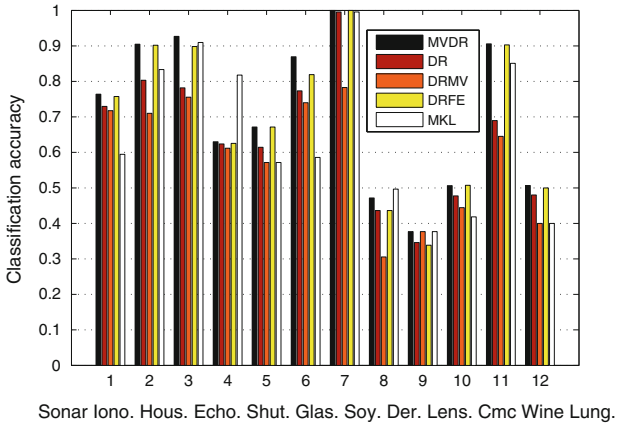


**Fig. 2** Classification accuracies of the algorithms: MVDR, DR, DRMV, DRFE, MKL [34]

single DR in terms of classification. Compared with SVM with RBF kernels, the proposed MVDR succeeds in some datasets (Sonar, Hous., Shut., Soy., Wine, Lung.) but fails in some datasets (Iono., Echo., Glas., Der., Lens., Cmc). Thus, our future work is to extend our method into the SVM framework.

Figure 2 shows the classification accuracies of these implemented algorithms on the data sets that are Sonar, Echocardiogram, Ionosphere, House-votes, Shuttle-landing-control, Glass, Soybean-small, Dermatology, Lenses, Cmc, Wine, Lung-cancer (denoted for short as Sonar, Iono., Hous., Echo., Shut., Glas., Soy., Der., Lens., Cmc, Wine, Lung., respectively). Figure 2 gives the histogram of the classification results. The higher the histogram is, the better its corresponding algorithm is. Then, it can be found that: (1) the proposed MVDR is superior to DR on all the used data sets; (2) the DRFE or the MKL [34] learning take the second or third place, and both are worse performance than the proposed MVDR in most cases.

In addition to reporting the average classification accuracies, we also perform the paired $t$-test [25] by comparing the proposed MVDR with the other algorithms DR, DRFE, DRMV and MKL [34]. The null hypothesis $H_0$ demonstrates that there is no significant difference between the mean number of the samples correctly classified by the proposed method and

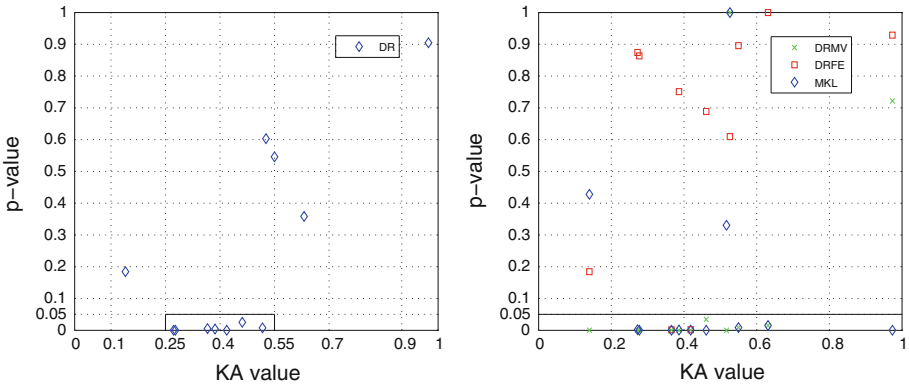


**Fig. 3** The $p$-value as a function of the kernel alignment (KA) value on the used data sets

**Table 3** Running time (in second) comparison between the algorithms MVDR, DR, DRMV, DRFE, MKL [34]

| Data sets | MVDR | DR | DRMV | DRFE | MKL |
|---|---|---|---|---|---|
| Sonar | 6.643 | 0.249 | 0.0749 | 0.5907 | 221.528 |
| Echo. | 0.1687 | 0.0124 | 0.038 | 0.1265 | 0.0953 |
| Iono. | 3.1548 | 0.145 | 0.6517 | 2.3487 | 68941.02 |
| Hous. | 5.7984 | 0.1406 | 0.6424 | 5.1593 | 36955.45 |
| Shut. | 0.0048 | 0.0015 | 0.0046 | 0.0015 | 0.0513 |
| Glas. | 0.0844 | 0.0438 | 0.0596 | 0.0656 | 2967.32 |
| Soy. | 0.0313 | 0.0015 | 0.0048 | 0.0139 | 0.2562 |
| Der. | 1.4252 | 0.1282 | 0.3892 | 1.5953 | 1.1874 |
| Lens. | 0.0015 | 0.0015 | 0.0016 | 0.0047 | 0.0406 |
| Cmc | 82.7547 | 21.9451 | 37.3142 | 136.1033 | 8.531 |
| Wine | 0.1124 | 0.000413 | 0.0265 | 0.0672 | 0.2045 |
| Lung. | 0.7283 | 0.00020 | 0.0545 | 0.614 | 0.147 |

the other algorithms. Under this assumption, the $p$-value of each test is the probability of a significant difference in the correctness values occurring between the two testing sets. Thus, the smaller the $p$-value, the less likely that the observed difference results from identical testing set correctness distributions. The threshold for the $p$-value is set to 0.05. Figure 3 gives all the $p$-values of the compared algorithms on the used data sets. From this figure, it can be found that: (1) the null hypothesis $H_0$ is rejected between MVDR and DR on 7 data sets, i.e., MVDR is significantly better than DR on these data sets; (2) except DRFE, $H_0$ is also rejected between the proposed method and DRMV, MKL [34] on most data sets used here.

### *5.2.2 Running Time*

Table 3 reports the training time of the proposed MVDR and those compared algorithms (DR, DRMV, DRFE and MKL [34]) with their optimal parameters in 10 runs. All the computations are performed on Pentium IV 2.80 GHz processor running Windows 2000 Terminal and MATLAB environment. From Table 3, although the proposed MVDR has a longer running time than DR on most of the data sets due to multiple kernels used, the proposed method has a significantly shorter running time with respect to the MKL [34] on most cases. Further,

compared with both DRMV and DRFE, it can also be noted that the proposed MVDR has a competitive efficiency.

### 5.3 Further Analysis of Multiple Views

The existing MVL such as co-training requires the conditional independence assumption well satisfied where the patterns are obtained from multiple sources [5]. However, Wang and Zhou [43] give a deep discussion on co-training style algorithms in semi-supervised problems, and theoretically demonstrate that the base learners with enough differences can lead to a superior performance in co-training style algorithms. They explain why co-training algorithms can succeed in some cases without two views. This paper extends the work of Wang and Zhou [43] and gets a similar conclusion on supervised problems. In the proposed algorithm MVDR, on the one hand only the single-view patterns are available. On the other hand, the generated views are induced from the multiple empirical kernel mappings. Thus we adopt kernel alignment [9] as a good correlation measure between the induced $M$ views to explore the reasons why the performance of the proposed MVDR can be improved. The definition of kernel alignment for two views is given as follows:

**Definition** (*Kernel Alignment* [9]) The alignment between the Gram matrices $K_i$ and $K_j$ (one empirical kernel can correspond to one Gram matrix) is

$$A_{ij} = \frac{tr\left(K_i^T K_j\right)}{\sqrt{tr\left(K_i^T K_j\right) tr\left(K_i^T K_j\right)}}. \tag{25}$$

Then the alignment between $M (M \geq 2)$ views is given as

$$A = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i}^{M} A_{ij}. \tag{26}$$

The $A$ value can be taken as the cosine value of the angle between the Gram matrices, it satisfies $-1 \leq A \leq 1$. Here, since the Gram matrix $K$ is positive semi-definite, $0 \leq A \leq 1$. Intuitively, the bigger the value of $A$, the more correlated the matrices and also the more correlated the corresponding views. If $A_{ij} = 1$, $K_i = \xi K_j, \xi \in \mathbb{R}$.

One '$\diamond$' ('$\times$' or '$\square$') in Fig. 3 denotes on *one certain data set*, what the $p$-value between MVDR and one certain algorithm is, and what its corresponding $A$ value of MVDR is. From the left sub-plot of Fig. 3, it can be clearly found that the $A$ values of those points ($p$-value $<0.05$) are most in the range from 0.25 to 0.55. In other words, the weaker correlation between the views leads to the performance improvement in the proposed MVDR. The similar result can also be found in the right sub-figure of Fig. 3. A further work about the relationship between the kernel alignment and MVDR will be implemented in future.

## 6 Conclusion

The contribution of this paper is to develop a novel MVL named MVDR on single-view patterns. The proposed MVDR maps the original single-view patterns into multiple feature spaces with different empirical kernels and associates each generated space with our previous work of DR, where the DR learning in each space is taken as one view of the proposed

MVDR. Simultaneously, the proposed MVDR has an analytic solution to the optimization problem and a direct optimization formulation for multi-class problems without one-against-all or one-against-one strategies. The experimental results show that the proposed method provides a complementarity between different views and thus has a superior classification performance to the original single-view algorithm DR. Further, compared with the other algorithms DRFE, DRMV and MKL [34], the proposed method has a better or competitive performance in terms of classification and computation. Finally, it is also found that the improved classification performance of our method is induced by a weak correlation between the views, which is validated by the experiments here.

**Acknowledgements** The authors thank Natural Science Foundations of China under Grant Nos. 60905002, 60903091, 60675027 and 60773061, Natural Science Foundations of Jiangsu Province Grant No. BK2008381, the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20060251013, and the High-Tech Development Program of China (863) under Grant No. 2006AA10Z315 for support. This work is also supported by the Open Projects Program of National Laboratory of Pattern Recognition.

## References

1. Bach F, Lanckriet GRG, Jordan MI (2004) Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st international conference on machine learning
2. Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 1:1–48
3. Bennett KP, Momma M, Embrechts MJ (2002) MARK: a boosting algorithm for heterogeneous kernel models. In: SIGKDD, pp 24–31
4. Bi J, Zhang T, Bennett K (2004) Column-generation boosting methods for mixture of kernels. In: KDD, pp 521–526
5. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proceedings of the conference on computational learning theory
6. Chapelle O, Vapnik V, Bousquet O, Mukherjee S (2002) Choosing multiple parameters for support vector machines. Mach Learn 46(1–3):131–159
7. Chen Z, Haykin S (2002) On different facets of regularization theory. Neural Comput 14(12):2791–2846
8. Chen S, Hong X, Harris C (2004) Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization. IEEE Trans Syst Man Cybern B 34(4):1708–1717
9. Cristianini N, Elisseef A, Shawe-Taylor J (2001) On kernel-target alignment. In: Advances in neural information processing systems
10. Dai D, Yuen P (2007) Face recognition by regularized discriminant analysis. IEEE Trans Syst Man Cybern B 37(4):1080–1085
11. de Diego IM, Moguerza JM, Munoz A (2004) Combining kernel information for support vector classification. In: MCS, LNCS, pp 102–111
12. Duda R, Hart P, Stork D (2001) Pattern classification. Wiley, New York
13. Duin R, Pekalska E (2006) Object representation, sample size and data complexity. In: Basu M, Ho TK (eds) Data complexity in pattern recognition. Springer, London, pp 25–47
14. Evgeniou T, Micchelli C, Pontil M (2005) Learning multiple tasks with kernel methods. J Mach Learn Res 6:615–637
15. Farquhar J, Hardoon D, Meng H, Shawe-Taylor J, Szedmak S (2005) Two view learning: SVM-2K, theory and practice. In: NIPS
16. Grandvalet Y, Canu S (2002) Adaptive scaling for feature selection in SVMs. In: Neural information processing systems
17. Guo P, Lyu M, Chen C (2003) Regularization parameter estimation for feedforward neural networks. IEEE Trans Syst Man Cybern B 33(1):35–44
18. Hardoon D, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16:2639–2664
19. Haykin S (2001) Neural networks: a comprehensive foundation. Tsinghua University Press, Beijing
20. Lanckriet GRG, Bie TD, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. Bioinformatics 20(16):2626–2635

21. Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. J Mach Learn Res 5:27–72
22. Lauer F, Bloch G (2007) Incorporating prior knowledge in support vector machines for classification: a review. Neurocomputing 71:1578–1594
23. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximun margin criterion. IEEE Trans Neural Netw 17(1):157–165
24. Martinez A, Kak A (2001) Pca versus lda. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
25. Mitchell TM (1997) Machine learning. McGraw-Hill, Boston
26. Momma M, Bennett K (2002) A pattern search method for model selection of support vector regression. In: Proceedings of the second SIAM international conference on data mining, SIAM, pp 261–274
27. Morozov V (1984) Methods for solving incorrectly posed problems. Springer, New York
28. Muslea I, Kloblock C, Minton S (2002) Active + semi-supervised learning = robust multi-view learning. In: ICML
29. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) Uci repository of machine learning databases. Available from: http://www.ics.uci.edu/mlearn/MLRepository.html
30. Nigam K, Ghani R (2000) Analyzing the effectiveness and applicability of co-training. In: Proceedings of information and knowledge management
31. Ong CS, Smola AJ, Williamson RC (2005) Learning the kernel with hyperkernels. J Mach Learn Res 6:1043–1071
32. Poggio T, Girosi F (1990) Regularization algorithms for learning that are equivalent to multilayer networks. Science 247:978–982
33. Poggio T, Smale S (2003) The mathematics of learning: dealing with data. Notices AMS 50(5):537–544
34. Rakotomamonjy A, Bach F, Canu S, Grandvalet Y (2007) More efficiency in multiple kernel learning. In: ICML
35. Scholkopf B, Mika S, Burges CJC, Knirsch P, Muller K-R, Ratsch G, Smola AJ (1999) Input space versus feature space in kernel-based methods. IEEE Trans Neural Netw 10(5):1000–1017
36. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University, Cambridge
37. Sonnenburg S, Ratsch G, Schafer C (2005) A general and efficient multiple kernel learning algorithm. In: Neural information processing systems
38. Szedmak S, Shawe-Taylor J (2005) Muticlass learning at one-class complexity. Technical report no: 1508, School of Electronics and Computer Science, Southampton, UK
39. Tikhonov A (1963) On solving incorrectly posed problems and method of regularization. Doklady Akademii Nauk USSR 151:501–504
40. Tikhonov A, Aresnin V (1977) Solutions of ill-posed problems. Winston, Washington, DC
41. Tsang I, Kocsor A, Kwok J (2006) Efficient kernel feature extraction for massive data sets. In: International conference on knowledge discovery and data mining
42. Vapnik V (1998) Statistical learning theory. Wiley, New York
43. Wang W, Zhou Z (2007) Analyzing co-training style algorithms. In: Proceedings of the 18th European conference on machine learning (ECML'07)
44. Wang Z, Chen S, Sun T (2008) Multik-MHKS: a novel multiple kernel learning algorithm. IEEE Trans Pattern Anal Mach Intell 30:348–353
45. Xiong H, Swamy MNS, Ahmad MO (2005) Optimizing the kernel in the empirical feature space. IEEE Trans Neural Netw 16(2):460–474
46. Xu QS, Liang YZ (2001) Monte carlo cross validation. Chemom Intell Lab Syst 56:1–11
47. Xue H, Chen S, Yang Q (2009) Discriminatively regularized least-squares classification. Pattern Recognit 42:93–104
48. Zhang P, Peng J (2004) SVM vs regularized least squares classification. In: Proceedings of the 17th international conference on pattern recognition
49. Zhang K, Tang J, Li J, Wang K (2005) Feature-correlation based multi-view detection. In: ICCSA 2005, LNCS 3483, pp 1222–1230
50. Zhou Y, Goldman S (2004) Democratic co-learning. In: Proceedings of the 16th IEEE international conference on tools with artificial intelligence (ICTAI2004)