Expert Systems with Applications 38 (2011) 11193–11200

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel multi-view classifier based on Nyström approximation

Zhe Wang [a,b], Songcan Chen [b,*], Daqi Gao [a]

[a] Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai 200237, PR China
[b] Department of Computer Science & Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China

ARTICLE INFO

Keywords:
Multi-view learning
Single-view patterns
Kernel-based method
Nyström approximations
Rademacher complexity
Classifier design

ABSTRACT

The existing multi-view learning (MVL) is learning from patterns with multiple information sources and has been proven its superior generalization to the conventional single-view learning (SVL). However, in most real-world cases, researchers just have single source patterns available in which the existing MVL is uneasily directly applied. The purpose of this paper is to solve this problem and develop a novel kernel-based MVL technique for single source patterns. In practice, we first generate different Nyström approximation matrices K_ps for the gram matrix G of the given single source patterns. Then, we regard the learning on each generated Nyström approximation matrix K_p as one view. Finally, different views on K_ps are synthesized into a novel multi-view classifier. In doing so, the proposed algorithm as a MVL machine can directly work on single source patterns and simultaneously achieve: (1) low-cost learning; (2) effectiveness; (3) the same Rademacher complexity as the single-view KMHKS; (4) ease of extension to any other kernel-based learning algorithms.

1. Introduction

The pattern is the handled object of the classifier and it is meaningful to consider the correlated knowledge of patterns in designing classifiers (Duin & Pekalska, 2006). In practice, patterns can be obtained through single or multiple information sources. If we regard each information source as one view of patterns, the kinds of patterns are sorted into single-view patterns and multi-view patterns. Correspondingly, the learning based on single-view and multi-view patterns can be sorted into single-view and multi-view learning (SVL and MVL), respectively. In Blum and Mitchell (1998), it has been demonstrated that the existing MVL approach has a significantly superior generalization ability to the conventional SVL.

Different from the existing MVL framework, in this paper we develop a novel kernel-based multi-view classifier whose underlying motivations are:

- It is known that the existing MVL requires the independence assumption that multiple views of patterns are independent given the class label (Blum & Mitchell, 1998; Muslea, Kloblock, & Minton, 2002). The independence assumption tries to guarantee a high performance to be achieved since it is unlikely for compatible classifiers trained on independent views to agree on an incorrect label. However, in most real-world applications, the independence assumption is hard to given since there are only one single-view pattern set available. In that case, there is not any natural way to partition the feature space (Duin & Pekalska, 2006; Muslea et al., 2002; Nigam & Ghani, 2000) and thus the existing MVL framework can not effectively work. The fact motivates us to develop a new MVL that can potentially create multiple views from a group of single-view patterns and then learn from multiple views simultaneously.
- In the existing MVL framework, the patterns are represented by multiple independent sets of attributes. Meanwhile, the base learners have the same architecture in each view and iteratively bootstrap each other. Here, we expect to utilize the multi-view technique due to its well-known superior generalization to the usual SVL. However, different from the existing learning on multi-view patterns, we adopt a new multi-view viewpoint. In the new multi-view viewpoint, a multi-view classifier is designed for the original single-view patterns. In doing so, the advantage of the proposed MVL can be inherited.

It can be found that the proposed multi-view classifier is novel but off-the-shelf. We handily modify the existing technique for a low-cost design. Here, our work falls into the kernel-based learning framework. Kernel learning plays an important role in many applications. In kernel-based leaning, a canonical algorithm is reformulated in terms of gram matrix G and all the prior information or knowledge of patterns is contained in G (Shawe-Taylor & Cristianini, 2004). But a large scale data set induces a $G \in \mathbb{R}^{n\times n}$ with large storages and causes a much computational complexity. To this end, the Nyström approximation to G (Williams & Seeger, 2001) was presented (Fowlkes, Belongie, Chung, & Malik, 2004; Kumar, Mohri, & Talwalkar,

* Corresponding author.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.166

2009; Williams & Seeger, 2001). Through choosing a set of rows and columns from the original G, the Nyström approximation can speed up kernel-based machines but without a significant decrease in the classification accuracy. In the proposed MVL framework, we regard each Nyström approximation to G as one view and therefore achieve multiple views from the original single-view patterns. Concretely speaking, we randomly choose *different* sets of rows and columns from G so as to give rise to multiple approximating Gram matrices K_ps, $p = 1,\ldots,M$. Each K_p bases on the same given single-view patterns, but possesses its own structural description for the patterns due to the different choices. Then, the learning on each K_p can be taken as one view. Consequently, we fuse these different views in one joint rather than separate optimization process and form a new multi-view classifier.

- The proposed multi-view classifier can own a superior classification performance to its corresponding single-view version since here randomly selecting different reduced sets from G to form its different approximations K_ps is an optimal sampling strategy for minimizing a model variation. In the proposed method, we randomly select a set of rows and columns from the original gram matrix G so as to generate a Nyström approximation K_p. In doing so multiple times, we can give a set of K_ps. It has been stated that the presented method syncretizes multiple K_ps in one learning process. The designed leaning with multiple K_ps is supposed to be robust against the worst possible scenario and can reduce the model deviation. Here, the proposed learning acts a similar role like that of cross validation or Monte–Carlo sampling schemes (Duda, Hart, & Stock, 2001; Lee & Huang, 2007).
- The proposed *multiviewization* technique can be applied to any other kernel-based algorithms like the state-of-the-art *kernelization* technique applied to the linear algorithms since only the gram matrix G needs to be manipulated.

In our implementation, we take the kernel Ho-Kashyap classifier with regularization (KMHKS) (Leski, 2004) as a paradigm due to both its similar principle to support vector machines (SVM) (Vapnik, 1998) of maximizing the separation margin and superior generalization performance, and then develop the multi-view Nyström approximation KMHKS (MVNA-KMHKS). In Williams and Seeger (2001), it has been proven that the kernel-based machine with the Nyström approximation to G has the comparable generalization to that without. Here, the experimental results have demonstrated that even in the current so-simple construction for views, the proposed MVNA-KMHKS can not only outperforms the single-view classifier with the Nyström approximation (NA-KMHKS), but also gets ahead of the original one (KMHKS).

The rest of this paper is organized as follows. Section 2 discusses the related work on the multi-view learning. We formulate the problem setting and give the description about the proposed multi-view classifier in Section 3. Section 4 investigates the generalization risk bound of the proposed multi-view method and finds that the Rademacher complexity of the proposed method does not increase but is still kept the same as that of the single-view KMHKS. Meanwhile, this section discusses the relationship between the proposed MVL and ensemble learning. Following that, we report on our experimental results. Finally, conclusions are given.

2. Related work

One typical example of MVL is web-page classification (Blum & Mitchell, 1998), 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). In Blum and Mitchell (1998), Blum & Mitchell design a co-training algorithm on the labeled and unlabeled web pattern sets composed of the two naturally-split views. On the labeled web set two classifiers are incrementally built with the corresponding views, and on each cycle each classifier labels the unlabeled webs and picks those with the highest confidence into the labeled set. The process repeats until the terminated condition is satisfied. 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 (2000) experimentally explore the co-training algorithm with or without the independence assumption, demonstrate that the co-training algorithm with a natural split of the attributes outperforms the ones 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. (2002) incorporate active learning in co-EM, and present co-EMT that outperforms both co-training and co-EM and has a robustness in view-correlation cases to some extent. Recently, the literature (Wang & Zhou, 2007) demonstrates that the co-training style algorithms could success in the case that the two learners have enough difference without the the independence assumption.

Although co-EMT and co-EM has 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, Tang, Li, and Wang (2005) 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 Zhang et al. (2005), 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. SVM-2 K (Farquhar, Hardoon, Meng, & Shawe-Taylor, 2005) utilizes the multi-kernel trick on the single-view patterns where for the same pattern, the two views are generated through two feature projections ϕ_A and ϕ_B with their corresponding kernels k_A and k_B. Then the process that kernel canonical correlation analysis (KCCA) (Hardoon, Szedmak, & Shawe-Taylor, 2004) combined by SVM is done on the two generated views. However, due to SVM itself, SVM-2 K also suffers from similar problems as the scalability with the number of the patterns and time-consuming quadratic programming (QP). Rather than dealing with the single-view patterns themselves, democratic co-learning (Zhou & Goldman, 2004) runs different algorithms on the single-view patterns, whose motives are that different learning algorithms have different inductive biases and that better performance can be made by the voted majority. However, in democratic co-learning, how to select those learning algorithms to be fused is still a problem due to lack of a measurable selection criterion.

Compared with CCFP, SVM-2 K and democratic co-learning, the multi-view classifier proposed by us has the following advantages: (1) it does not need to split the attributes of the original single-view patterns but just manipulates the corresponding gram matrix; (2) it employs the off-the-shelf learning technique, i.e. the Nyström approximation to gram matrix, and simply generates multiple views naturally and freely.

3. The proposed multi-view classifier

3.1. Nyström approximation to gram matrix G

The Nyström approximation matrix $K \in \mathbb{R}^{n\times n}$ to gram matrix $G \in \mathbb{R}^{n\times n}$ (Williams & Seeger, 2001) is to decrease the computation cost of kernel-based methods from $O(n^3)$ to $O(m^2n)$, where n is the number of training patterns and $0 < m < n$. The approximation K is got by randomly choosing m rows and columns from G without replacement. Concretely, let G be partitioned into four blocks $G_{m,m}$, $G_{n-m,m} = G^T_{m,n-m}$ and $G_{n-m,n-m}$ and then set the approximating gram matrix

$$K = G_{n,m}G^{-1}_{m,m}G_{m,n}, \tag{1}$$

where

$$K_{m,m} = G_{m,m},$$
$$K_{n-m,m} = G_{n-m,m},$$
$$K_{m,n-m} = G_{m,n-m},$$
$$K_{n-m,n-m} = G_{n-m,m}G^{-1}_{m,m}G_{m,n-m}.$$

The experiments in Williams and Seeger (2001) demonstrate that when m is set to the empirical values 256 or 512 on the US Postal Service (USPS) handwritten digit database with 7291 training patterns (n = 7291), the classifier with the Nyström approximation has the comparable classification accuracy to the one without, i.e. the average number of errors on the test patterns are 26 vs. 26.1.

3.2. Multi-view Nyström approximation KMHKS (MVNA-KMHKS)

Suppose that there are n labeled training patterns $\{(x_i, y_i)\}^n_{i=1}$ available, where $x_i \in \mathbb{R}^d$ and the corresponding class label $y_i \in \{+1, -1\}$. The training set can give rise to a gram matrix G by a given kernel function $k(x_i,x_j)$ (Shawe-Taylor & Cristianini, 2004). In KMHKS (Leski, 2004), the decision function

$$f(x) = sign\left(\sum_{i=1}^{n} y_i\alpha_i k(x, x_i) + w_0\right), \tag{2}$$

is obtained by optimizing the criterion

$$\min J(\Gamma, \mathbf{b}, w_0) = (G\Gamma + w_0Y - \mathbf{1} - \mathbf{b})^T(G\Gamma + w_0Y - \mathbf{1} - \mathbf{b}) + c\Gamma^TG\Gamma, \tag{3}$$

where the scalars $\alpha_i, w_0 \in \mathbb{R}$, $c \geqslant 0$; the vectors $\Gamma_{n\times1} = [\alpha_i]^n_{i=1}$, $Y_{n\times1} = [y_i]^n_{i=1}$, $\mathbf{1}_{n\times1}$ and $\mathbf{b}_{n\times1}$ respectively denote the vectors of dimension $n \times 1$ with all entries equal to 1 and the nonnegative value; the gram matrix $G = [y_iy_jk(x_i, x_j)]^n_{i,j=1}$.

In our method, through the random choices for m (different sets of rows and columns) from the gram matrix G, there are multiple Nyström approximating gram matrices K_ps, $p = 1\ldots M$. Each K_p is associated with one KMHKS, which learns from the same patterns but owns different structural architecture (view) due to the different approximation. Consequently, it induces multiple views KMHKSs. In such a set of views, there is a set of solutions $\{\Gamma^p, w^p_0\}^M_{p=1}$, correspondingly. A natural idea is to learn the solution set $\{\Gamma^p, w^p_0\}^M_{p=1}$ such that each individual KMHKS respectively on its corresponding approximating matrix can correctly classify a given pattern. It is known that although the given pattern generates multiple views through different Nyström approximations, it still has one class label. In other words, the disagreements among the outputs of all KMHKSs should be farthest minimized. This suggests the following objective function of multi-view Nyström approximation KMHKS (MVNA-KMHKS)

$$\begin{aligned}\min J'(\Gamma^p, \mathbf{b}^p, w^p_0) = &\sum_{p=1}^{M}\Big((K_p\Gamma^p + w^p_0Y - \mathbf{1} - \mathbf{b}^p)^T(K_p\Gamma^p \\ &+ w^p_0Y - \mathbf{1} - \mathbf{b}^p) + c^p\Gamma^{pT}K_p\Gamma^p\Big) \\ &+ \gamma\sum_{p=1}^{M}\left(K_p\Gamma^p + w^p_0Y - \sum_{j=1}^{M}\mu_j(K_j\Gamma^j + w^j_0Y)\right)^T \\ &\times\left(K_p\Gamma^p + w^p_0Y - \sum_{j=1}^{M}\mu_j(K_j\Gamma^j + w^j_0Y)\right),\end{aligned} \tag{4}$$

where $\mathbf{b}^p$, c^p are respectively the error vector and the regularization parameter of each view, γ is the coupling parameter that regularizes multiple views towards the compatibility using the multiple approximating gram matrices $\{K_p\}^M_{p=1}$ on the given single-view patterns, $\mu_j \geqslant 0$, $\sum^M_{j=1}\mu_j = 1$, μ_j denotes the importance of the corresponding view and the bigger the μ_j is, the more important the corresponding view is. The first term of the right side of (4) is to guarantee each view can correctly classify the patterns, and the second one is to minimize the disagreement between each view by making the output of each view be maximally close to the weight average output of all views.

By differentiating (4) with respect to Γ^p, w^p_0 and zeroing them, we obtain

$$\begin{aligned}&\Big((1+\gamma)K^T_pK_p + c^pK_p\Big)\Gamma^p + (1+\gamma)K^T_pYw^p_0 \\ &= K^T_p\left(\mathbf{1} + \mathbf{b}^p + \gamma\sum_{j=1}^{M}\mu_j\Big(K_j\Gamma^j + w^j_0Y\Big)\right),\end{aligned} \tag{5}$$

$$\begin{aligned}&(1+\gamma)Y^TK_p\Gamma^p + (1+\gamma)Y^TYw^p_0 \\ &= Y^T\left(\mathbf{1} + \mathbf{b}^p + \gamma\sum_{j=1}^{M}\mu_j\Big(K_j\Gamma^j + w^j_0Y\Big)\right).\end{aligned} \tag{6}$$

Then, considering that K_p is positive semi-definite and defining the matrix

$$\Lambda = \begin{bmatrix}(1+\gamma)K_p + c^pI & (1+\gamma)Y \\ (1+\gamma)Y^TK_p & (1+\gamma)Y^TY\end{bmatrix},$$

(5) and (6) are converted into

$$\begin{aligned}&\begin{bmatrix}\Gamma^p_{t+1} \\ w^p_{0t+1}\end{bmatrix} \\ &= \Lambda^{-1}\begin{bmatrix}\mathbf{1} + \mathbf{b}^p_t + \gamma\left(\sum_{j=1}^{p-1}\mu_j\Big(K_j\Gamma^j_{t+1} + w^j_{0t+1}Y\Big) + \sum_{j=p}^{M}\mu_j\Big(K_j\Gamma^j_t + w^j_{0t}Y\Big)\right) \\ Y^T\left(\mathbf{1} + \mathbf{b}^p_t + \gamma\left(\sum_{j=1}^{p-1}\mu_j\Big(K_j\Gamma^j_{t+1} + w^j_{0t+1}Y\Big) + \sum_{j=p}^{M}\mu_j\Big(K_j\Gamma^j_t + w^j_{0t}Y\Big)\right)\right)\end{bmatrix},\end{aligned} \tag{7}$$

where the subscript t denotes the iteration index. The gradient of (4) with respect to $\mathbf{b}^p$ is given as follows

$$\nabla_{\mathbf{b}^p}J' = -2\big(K_p\Gamma^p + w^p_0Y - \mathbf{1} - \mathbf{b}^p\big). \tag{8}$$

In order to keep the condition $\mathbf{b}^p \geqslant 0$ in each view, we start with $\mathbf{b}^p_1 \geqslant 0$, refuse to decrease any of its components like KMHKS, and give the update of $\mathbf{b}^p$ as follows

$$\begin{cases}\mathbf{b}^p_1 > 0 \\ \mathbf{b}^p_{t+1} = \mathbf{b}^p_t + \rho^p(\mathbf{e}^p_t + |\mathbf{e}^p_t|)\end{cases}, \tag{9}$$

where at the tth iteration, the error vector of the pth view $\mathbf{e}^p_t = K_p\Gamma^p_t + w^p_{0t}Y - \mathbf{1} - \mathbf{b}^p_t$, and the learning rate of the pth view $0 < \rho^p < 1$. In practice, the termination criterion can be designed as

$$\frac{\|J'_{t+1} - J'_t\|}{\|J'_t\|} \leqslant \xi, \tag{10}$$

Table 1
Algorithm **MVNA-KMHKS**.

Input: The single-view patterns $\{(x_i, y_i)\}_{i=1}^n$;
M approximating gram matrices $\{K_p\}_{p=1}^M$.
OutPut: $\{\Gamma^p, w_0^p\}_{p=1}^M$.
1. Initialize $\Gamma_1^p,\ w_{01}^p,\ \mathbf{b}_1^p \geqslant 0,\ p = 1,\ldots,M$ at random; LET $t = 1$;
2. Do until the termination criterion (10) is satisfied:
   (A) For $p = 1,\ldots,M$:
      I. Compute $\Gamma_{t+1}^p, w_{0t+1}^p$ with (7);
      II. Set $\mathbf{b}_{t+1}^p$ with (9);
   (B) Compute J_{t+1}^t with (4);
   (C) Increment t.
3. Return the final $\{\Gamma^p, w_0^p\}_{p=1}^M$.

where $\xi \in \mathbb{R}$ is a small positive value, and $\|.\|$ is chosen to be L_2 norm throughout the paper. Such a designed procedure is exactly the MVNA-KMHKS and summarized in Table 1.

The decision function of MVNA-KMHKS for the pattern $x \in \mathbb{R}^d$ is given as follows

$$g(x) = sign\left(\sum_{p=1}^{M}\sum_{i=1}^{n}\mu_p\left(y_i\alpha_i^p k(x,x_i) + w_0^p\right)\right), \tag{11}$$

where $\Gamma^p = [\alpha_i^p]_{i=1}^n$.

It can be found that in Algorithm MVNA-KMHKS, the update of $\Gamma_{t+1}^p,\ w_{0t+1}^p$ is determined by $\{\Gamma_{t+1}^j, w_{0t+1}^j\}_{j=1}^{p-1}$ and $\{\Gamma_t^j, w_{0t}^j\}_{j=p}^M$ as in (7), which reflects that these views *cooperate* each other. Then, if $M = 1$, $\gamma = 0$ of (4), MVNA-KMHKS will be degenerated to KMHKS and thus KMHKS is taken as a special instance of MVNA-KMHKS.

4. Discussion

4.1. Rademacher complexity analysis

It is well-known that the analysis of the generalization risk bound is important for algorithms (Bartlett, Boucheron, & Lugosi, 2002; Koltchinskii & Panchenko, 2000; Koltchinskii, 2001; Vapnik & Chervonenkis, 1971). For example, the generalization risk bound can be used to choose a suitable model. In this section, we give the discussion of the proposed MVNA-KMHKS in terms of the generalization risk bound with the Rademacher complexity. Firstly, we know that the classical risk bound theory was proposed by Vapnik and Chervonenkis (1971) and can be described through Theorem 1.

Theorem 1. *Let P be a probability distribution on $\chi \times \{\pm 1\}$ and $\{x_i, y_i\}_{i=1}^n$ be chosen independently according to P. Then, for a $\{\pm 1\}$-valued function class F with the domain χ, there is a constant $c \geqslant 0$ such that for any integer n, with probability at least $1-\delta$ over $\{x_i, y_i\}_{i=1}^n$, every f in F satisfies*

$$P(y \neq f(x)) \leqslant \hat{P}_n(y \neq f(x)) + c\sqrt{\frac{VC(F)}{n}}, \tag{12}$$

where $VC(F)$ denotes the Vapnik–Chervonekis dimension of F and $\hat{P}_n$ denotes the empirical risk error of the function f on the sample set $\{x_i, y_i\}_{i=1}^n$.

In this case, the VC (F) dimension measures the complexity of the class function F. Further, the Rademacher complexity was proposed as an alternative notion of the complexity of a function class F (Bartlett & Mendelson, 2002; Farquhar et al., 2005; Koltchinskii, 2001; Mendelson, 2002). Here, the Rademacher complexity is used to measure the proposed MVNA-KMHKS. Definition 2 gives the Rademacher complexity (Koltchinskii, 2001).

Definition 2. Let μ be a probability distribution on a set χ and suppose that $\{x_i\}_{i=1}^n$ are independent samples selected from χ according to μ. Let F be a class of functions mapping from χ to $\mathbb{R}$. Let $\{\sigma_i\}_{i=1}^n$ be independent uniform $\{\pm 1\}$-valued random variables and define the random variable

$$\widehat{R}_n(F) = \mathbf{E}[\sup_{f\in F}|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\|x_1,\ldots,x_n], \tag{13}$$

where $\mathbf{E}$ is the operator of the expected value of a random variable. Then the Rademacher complexity of F is

$$R_n(F)F = \mathbf{E}\widehat{R}_n(F). \tag{14}$$

Theorem 3 (Bartlett & Mendelson, 2002) gives the generalization risk bound of F with the Rademacher complexity $R_n(F)$.

Theorem 3. *Let P be a probability distribution on $\chi \times \{\pm 1\}$ and $\{x_i, y_i\}_{i=1}^n$ be chosen independently according to P. Then, for a $\{\pm 1\}$-valued function class F with the domain χ, with probability at least $1-\delta$ over $\{x_i, y_i\}_{i=1}^n$, every f in F satisfies*

$$P(y \neq f(x)) \leqslant \hat{P}_n(y \neq f(x)) + \frac{R_n(F)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{15}$$

Due to the Eqs. (2) and (11) with kernels, the proposed method MVNA-KMHKS falls into the kernel-based learning framework. According to the literature (Bartlett & Mendelson, 2002), the empirical Rademacher complexity of F (13) also satisfies

$$\widehat{R}_n(F) \leqslant \frac{2B}{n}\sqrt{\sum_{i=1}^{n}k(x_i,x_i)} \leqslant \frac{2B}{n}\sqrt{tr(G)}, \tag{16}$$

where G is the gram matrix formed by a given kernel function $k(x_i,x_j)$ on the sample set $\{x_i\}_{i=1}^n$, the parameter B is a fixed value and satisfies $\Gamma^T G\Gamma \leqslant B^2$, $\Gamma \in \mathbb{R}^n$.

Then, the Rademacher complexity of F (16) can also satisfy

$$R_n(F) = \mathbf{E}\widehat{R}_n(F) \leqslant 2B\sqrt{\frac{\mathbf{E}tr(G)}{n}}. \tag{17}$$

Therefore, Theorem 3 can also be rewritten into

Theorem 3′. *Let P be a probability distribution on $\chi \times \{\pm 1\}$ and $\{x_i, y_i\}_{i=1}^n$ be chosen independently according to P. Then, for a $\{\pm 1\}$-valued function class F with the domain χ, with probability at least $1-\delta$ over $\{x_i, y_i\}_{i=1}^n$, every f in F satisfies*

$$P(y \neq f(x)) \leqslant \hat{P}_n(y \neq f(x)) + B\sqrt{\frac{\mathbf{E}tr(G)}{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{18}$$

Here, the generalization risk bound of the single-view KMHKS satisfies the inequation (18). According to the Eqs. (2) and (11), the decision function g of the proposed multi-view MVNA-KMHKS is the convex combination of the decision function f of the single-view KMHKS. It has been proven that for a class of functions F, if $convF$ is the class of convex combinations of function from $F, -F = \{-f : f \in F\}$ (Bartlett & Mendelson, 2002) then

$$R_n(convF) = R_n(F). \tag{19}$$

Concretely, for the sample set $\{x_i\}_{i=1}^n$ and $\{\sigma_i\}_{i=1}^n$,

$$\begin{aligned}\sup_{g\in convF}|\sum_{i=1}^{n}\sigma_i g(x_i)| &= \max\left(\sup_{g\in convF}\sum_{i=1}^{n}\sigma_i g(x_i), \sup_{g\in convF}-\sum_{i=1}^{n}\sigma_i g(x_i)\right)\\ &= \max\left(\sup_{f\in F}\sum_{i=1}^{n}\sigma_i f(x_i), \sup_{f\in F}-\sum_{i=1}^{n}\sigma_i f(x_i)\right)\\ &= \sup_{f\in F}|\sum_{i=1}^{n}\sigma_i f(x_i)|.\end{aligned}$$

Further, according to the definition of the Rademacher complexity, the Eq. (19) is proven (Bartlett & Mendelson, 2002).

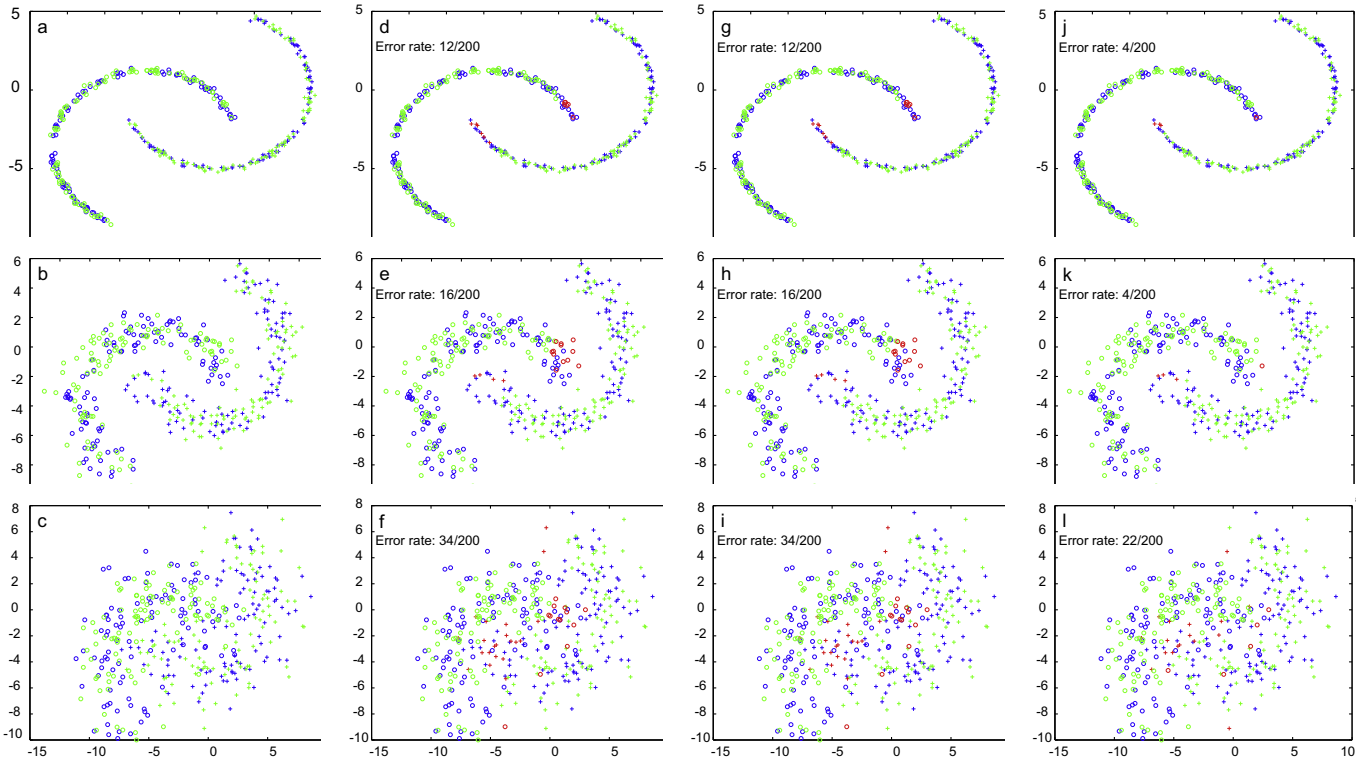

Fig. 1. Two-Moon Data: KMHKS (d, e, f), NA-KMHKS (g, h, i) and MVNA-KMHKS (j, k, l) (Denote the 3 subfigures in the 1st column as a, b and c and similarly, the rest subfigures in the 2nd, 3rd, and 4th columns are denoted from d to l, respectively.)

5.3. UCI data

5.3.1. Classification performance

To further investigate the classification performance of the proposed MVNA-KMHKS, it is compared with its single-view version NA-KMHKS, the original KMHKS and the state-of-the-art SVM on some UCI data sets where the one-against-one classification strategy (Krebel, 1998) is adopted for multi-class problems. For each data set, their classification accuracies on the testing sets generated by the 10-folds cross validation are averaged and reported in Table 2, where for the different methods, the best results are in bold. In addition to reporting the average accuracies, we perform the paired *t*-test (Mitchell, 1997) by comparing MVNA-KMHKS with the other classifiers NA-KMHKS, KMHKS, and SVM. The null hypothesis H_0 demonstrates that there is no significant difference between the mean number of samples correctly classified by MVNA-KMHKS and the other classifiers. Under this assumption, the *p*-value of each test is the probability of a significant difference in correctness values occurring between 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 *p*-value is set to 0.05. From this table, it can be found that: (1) the average classification accuracy of MVNA-KMHKS is superior to that of the other classifiers NA-KMHKS, KMHKS, and SVM on all the used data sets only except Iris; (2) NA-KMHKS has the comparable performance to KMHKS on House-votes, Horse-colic, Echocardiogram, and Iris, but clearly failed on Lung-cancer and Shuttle-landing-control; (3) on both Lung-cancer and Shuttle-landing-control, MVNA-KMHKS succeeds and even clearly outperforms KMHKS; (4) further, based on the *p*-value MVNA-KMHKS has significantly different accuracies from others on Lung-cancer and Horse-colic.

Table 2
Average testing accuracy (%) and *p*-values comparison between MVNA-KMHKS, NA-KMHKS, KMHKS and SVM (Note: The best testing results of each data set are in bold. The *p*-values are from a *t*-test comparing each classifier to MVNA-KMHKS. An asterisk * denotes that the difference from MVNA-KMHKS is significant at 5% significance level, i.e. *p*-value less than 0.05. The kernel alignment (KA) values of MVNA-KMHKS on each set are given in italic.)

Data set	MVNA-KMHKS Accuracy *KA-value*	NA-KMHKS Accuracy *p*-value	KMHKS Accuracy *p*-value	SVM Accuracy *p*-value
Lung-cancer	**57.33**	46.67*	50.00	52.67
	0.5169	0.0107	0.1116	0.3153
House-votes	**93.98**	93.44	93.94	93.80
	0.3261	0.1990	0.9275	0.7414
Shuttle-landing-control	**71.43**	62.86	70.00	70.00
	0.2790	0.1233	0.8084	0.7946
Horse-colic	**63.46**	60.63*	60.63*	59.69*
	0.4249	0.0478	0.0478	0.0043
Echocardiogram	**88.51**	87.31	87.61	88.36
	0.0759	0.3915	0.5228	0.9038
Iris	97.33	97.07	97.07	**97.60**
	0.9801	0.6750	0.7006	0.6111

The best testing results of the compared methods on each data set are in bold. The kernel alignment (*KA*) values of MVNA-KMHKS for each set are given in italic and shown in the second column.

5.3.2. Correlation in multiple views

The existing MVL such as co-training requires the independence asumption well satisfied (Blum & Mitchell, 1998) where the patterns are obtained from multiple sources. In our method, on the one hand only the single-view patterns are available. On the other hand, the views are induced from the multiple approximating gram matrices and the number of the views is set to 2 ($M = 2$). Thus we adopt the kernel alignment (KA) (Cristianini et al., 2001) as a good correlation measure between views to further explore the reasons why the performance of our method improves. Its definition is given as follows:

Table 3
Comparison of the error numbers of the MVNA-KMHKS, NA-KMHKS, and KMHKS for the 9 different tasks.

	0	2	3	4	5	6	7	8	9
MVNA-KMHKS	1	1	21	4	10	5	17	22	18
NA-KMHKS	2	10	25	5	14	5	19	23	21
KMHKS	2	8	26	4	11	6	19	28	18

Definition 3′ (***Kernel alignment*** *(Cristianini et al., 2001)*). The correlation between the gram matrices K_1 and K_2 is

$$A(K_1,K_2)=\frac{tr(K_1^TK_2)}{\sqrt{tr(K_1^TK_1)tr(K_2^TK_2)}}, \tag{21}$$

where tr(.) is a matrix trace operation.

The KA can be taken as the cosine of the angle between the gram matrices, it satisfies $-1\leqslant A(K_1,K_2)\leqslant 1$. Here, since K_i in (21) is substituted with the approximating gram matrix K_p that is positive semi-definite, $0\leqslant A(K_1,K_2)\leqslant 1$. Intuitively, the bigger the value of $A(K_1,K_2)$, the more correlated the matrices and also the more correlated the views from the matrices. If $A(K_1,K_2)=1$, $K_1=\alpha K_2,\ \alpha\in\mathbb{R}$. In our experiments, the KA values on each data set are also shown by italic in Table 2, from which it can be clearly obtained that the KA value only on Iris is nearly to 1 (0.9801) and all the others are no larger than 0.5169. The fact is consistent with the result that MVNA-KMHKS loses the first place only on Iris. Thus, the less correlation between the views of our method is the necessary condition of the performance improvement in the proposed multi-view classifier.

5.3.3. Further analysis in optical handwritten digits

Our method inherits the low computational cost of the Nyström approximation manipulating the gram matrix with $O(m^2n)$, and has the cost $O(Mm^2n)$ where M is the number of the generated views and is set to 2. So MVNA-KMHKS has the good scalability with the number of patterns. Here, the experiments were carried out on the relatively-large data set Optical Handwritten Digits (Optdigits) (Newman et al., 1998) that has 3823 training patterns and 1797 testing patterns. Due to the easy recognition on digit '1' and our limited condition on the computer and MATLAB environment, we select the digits '0,2,...,9' as the discussion example and m is set to 512 as used in Williams and Seeger (2001). Table 3 gives the number of errors for the three classifiers MVNA-KMHKS, NA-KMHKS, and KMHKS on the nine classification tasks. The results show that MVNA-KMHKS takes the first place in all the nine tasks especially for digit '2'.

Finally, we have also experimentally explored the convergence of the proposed MVNA-KMHKS. The results show that MVNA-KMHKS can converge in the limited iterations.

6. Conclusions

In this paper, our contributions mainly lie in.

- Significance: this paper introduces the creation of multiple views from a single view for multi-view learning. It is important because while the existing MVL has been shown to be effective, it relies heavily on the natural separability of the feature set into two independent components. In many settings, there might not be any natural way to partition the feature space, and the existing MVL framework may therefore not be applicable. In such scenarios, the proposed approach suggested in this paper can potentially create multiple independent or at least weaker correlated views from a single view and then learn from the multiple views simultaneously.
- Novelty in the two aspects: in the first aspect, the learning approach proposed in this paper is different from the existing multi-view learning approach. Instead of the classifiers trained on two different views iteratively boot-strapping each other, this paper proposes a joint learning approach that minimizes disagreement across the classifications using multiple views. There are similarities with ensemble learning, where predictions from different classifiers over a single view are combined, but, again, the critical difference is in the joint optimization. In the second aspect, for sake of comparison with the two current typical combined classifiers: Bagging or AdaBoost based on pattern sampling and Attribute Bagging based on attribute sampling both for generation of classifier diversity, our strategy is neither sampling patterns nor sampling attributes, instead constructing different views for our classifier from multiple approximating gram matrices induced through different Nyström approximations to the gram matrix. Each learning with the Nyström approximation matrix can develop a corresponding classifier and then can be combined together. As a result, a performance gain can be obtained. More importantly, we give an analysis of the generalization risk bound of the proposed MVNA-KMHKS and conclude that the Rademacher complexity of the proposed multi-view method does not increase but is still kept the same as that of the single-view KMHKS.

Acknowledgments

The authors thank Natural Science Foundations of China under Grant Nos. 60903091 and 61035003, the Specialized Research Fund for the Doctoral Program of Higher Education under Grant Nos. 200802870003 and 20090074120003 for support. This work is also partially supported by the Open Projects Program of National Laboratory of Pattern Recognition.

References

Bartlett, P., Boucheron, S., & Lugosi, G. (2002). Model selection and error estimation. *Machine Learning, 48*, 85–113.
Bartlett, P., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research, 3*, 463–482.
Basak, J., & Kothari, R. (2004). Classification paradigm for distributed vertically partitioned data. *Neural Computation, 16*(7), 1525–1544.
Blum, A., & Mitchell, T. (1998) Combining labeled and unlabeled data with co-training. In *Proceedings of the Conference on Computational Learning Theory*.
Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.
Cristianini, N., Elisseef, A., & Shawe-Taylor J. On kernel-target alignment. In *Advances in neural information processing systems*, 2001.
Duda, R., Hart, R., & Stock, D. (2001). *Pattern Classification* (2nd ed.). Wiley.
Duin, R., & Pekalska, E. (2006). Object representation sample size and data complexity. In M. Basu & T. K. Ho (Eds.), *Data complexity in pattern recognition* (pp. 25–47). London: Springer.
Farquhar, J., Hardoon, D., Meng, H., & Shawe-Taylor, J. (2005). Two view learning: Svm-2k, theory and practice. In *Neural Information Processing Systems*.
Fowlkes, C., Belongie Chung, F., & Malik, J. (2004). Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(2), 1–12.
Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceeding international conference on machine learning*.
Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119–139.
Hardoon, D., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation, 16*, 2639–2664.
Igelnik, B., Pao, Y., LeClair & Shen, C. (1999). The ensemble approach to neural-network learning and generalization. *IEEE Transactions on Neural Networks, 10*(1), 19–30.
Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory, 47*(5), 1902–1914.
Koltchinskii, V., & Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In E. Gine, D. Mason, & J. Wellner (Eds.), High dimensional probability II (pp. 443–459).

Krebel, U. (1998). Pairwise classification and support vector machines. In B. Scholkopf, C. Burges, & A. Somla (Eds.), *Advances in kernel methods: Support vector machine* (pp. 255–268). Cambridge, MA: MIT Press.

Kumar, S., Mohri, M., & Talwalkar, A. (2009). Ensemble nystrom method. In *Neural Information Processing Systems*.

Lee, Y., & Huang, S. (2007). Reduced support vector machines: A statistical theory. *IEEE TNN, 18*(1), 1–13.

Leski, J. (2004). Kernel ho-kashyap classifier with generalization control. *International Journal of Applied Mathematics and Computer Science, 14*(1), 53–61.

Mendelson, S. (2002). Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory, 48*(1), 251–263.

Mitchell, T. M. (1997). *Machine learning*. Boston: McGraw-Hill.

Muslea, I., Kloblock, C., & Minton, S. (2002) Active+semi-supervised learning = robust multi-view learning. In *Proceedings of the international conference on machine learning*.

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). Uci repository of machine learning databases. Available from: <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Nigam, K., & Ghani, R. (2000) Analyzing the effectiveness and applicability of co-training. In *Proceedings of information and knowledge management*.

Robert, B., Ricardo, G. O., & Francis, Q. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition, 36*, 1291–1302.

Seewald, A. (2003). Towards a theoretical framework for ensemble classification. In *Proceedings of the eighteenth international joint conference on artificial intelligence (IJCAI'03)* (pp. 1443–1444).

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University.

Tsang, I., Kocsor, A., & Kwok, J. (2006). Efficient kernel feature extraction for massive data sets. In *International conference on knowledge discovery and data mining*.

Valentini, G., & Masulli, F. (2002). Ensembles of learning machines. In M. Marinaro & R. Tagliaferri (Eds.), *Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences*. Heidelberg (Germany): Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications, 2*, 264–280.

Wang, w., & Zhou, Z. (2007) Analyzing co-training style algorithms. In *Proceedings of the 18th european conference on machine learning (ECML'07)* (pp. 454–465).

Williams, C., & Seeger, M. (2001) Using the nystrom method to speed up kernel machines. In *Advances in neural information processing systems*.

Windeatt, T. (2006). Accuracy/diversity and ensemble mlp classifier design. *IEEE Transactions on Neural Networks, 17*(5), 1194–1211.

Zhang, K., Tang, J., Li, J., & Wang, K. (2005) Feature-correlation based multi-view detection. In *ICCSA 2005, LNCS, vol. 3483* (pp. 1222–1230).

Zhou, Y., & Goldman, S. (2004) Democratic co-learning. In *Proceedings of the 16th IEEE international conference on tools with artificial intelligence (ICTAI 2004)*.