An Iterative Fusion Approach to Graph-based Semi-supervised Learning from multiple views

Yang Wang[†‡], Jian Pei[§], Xuemin Lin[†], and Qing Zhang[‡†]

[†]The University of New South Wales, Sydney, Australia
{wangy,lxue}@cse.unsw.edu.au

[§]Simon Fraser University, Canada
jpei@cs.sfu.ca

[‡]Australian E-Health Research Center
qing.zhang@csiro.au

Abstract. Often, a data object described by many features can be naturally decomposed into multiple "views", where each view consists of a subset of features. For example, a video clip may have a video view and an audio view. Given a set of training data objects with multiple views, where some objects are labeled and the others are not, *semi-supervised learning with graphs from multi-views* tries to learn a classifier by treating each view as a similarity graph on all objects, where edges are defined by the similarity on object pairs based on the view attributes. Labels and label relevance ranking scores of labeled objects can be propagated from labeled objects to unlabeled objects on the similarity graphs so that similar objects receive similar labels. The state-of-the-art, one-combo-fits-all methods linearly and independently combine either the metrics or the label propagation results from multi-views and then build a model based on the combined results. However, more often than not, the similarities between various objects may be manifested differently by different views. In such situations, the one-combo-fits-all methods may not perform well. To tackle the problem, we develop an iterative Semi-Supervised Metric Fusion (SSMF) approach in this paper. SSMF fuses metrics and label propagation results from multi-views iteratively until the fused metric and label propagation results converge simultaneously. Views are weighted dynamically during the fusion process so that the adversary effect of irrelevant views, identified at each iteration of fusion process, can be reduced effectively. To evaluate the effectiveness of SSMF, we apply it on multi-view based and content based image retrieval and multi-view based multi-label image classification on real world data set, which demonstrates that our method outperforms the state-of-the-art methods.

1 Introduction

Semi-supervised learning with graphs [10] is an important and effective approach, which propagates limited label information to unlabeled data objects on a similarity graph. A similarity graph uses the set of objects as vertices, and links edges based on the similarity between objects. Edges in a similarity graph may take similarity scores as weights. After *label propagation* [10] or *manifold ranking* [9] in a similarity graph, the more similar two objects, the more likely they have similar labels or the similar label relevance ranking scores. This property

is called *local smoothness* [8]. The labeled objects iteratively propagate the label information or label relevance ranking scores to unlabeled ones via graph edges until convergence, and the final labeling result based on the label relevance scores should be consistent to the initial label information, which is called *global consistency* [8].

Often, a data object described by many features can be naturally decomposed into multiple "views", where each view consists of a subset of features. For example, an image may have a color view and a shape view. Given a set of training data objects with multi-views, where some objects are labeled and the others are not, *semi-supervised learning with graphs from multi-views* tries to learn a classifier by incorporating the complementary information from multi-views. The state-of-the-art methods conduct in a "one-combo-fits-all" manner. That is, they linearly and independently combine either the metrics or the label propagation (manifold ranking) results from multi-views and then build a model based on the combined results. Specifically, the *metric fusion first strategy* [2] obtains a linear fusion of the metrics from multi-views, constructs a single similarity graph based on the fused metric, and conducts label propagation or manifold ranking on the similarity graph. Alternatively, the *propagation fusion strategy* [3, 5, 6] conducts label propagation or manifold ranking on each view first, and then obtains a linear fusion of label propagation results based on label relevance ranking scores in multi-views as the overall label propagation results.

More often than not, the similarities between various objects may be manifested differently by different views. For example, two video clips that are the same advertisement video but in different languages have similar video content but very different audio content. At the same time, two video clips that are two advertisements from the same company for the same campaign on the same product may have similar audio content but different video content. In such situations, the one-combo-fits-all methods may not perform well, since they use the same linear fusion from multi-views for all objects. Moreover, different views in such methods don't collaborate with each other to achieve consistency when performing fusion process.

To tackle the problem, in this paper, we develop an iterative fusion approach, called SSMF (for semi-supervised metric fusion and cross-view label propagation). SSMF fuses metrics and label propagation results from multi-views iteratively until the fused metric and label propagation results converge simultaneously. Views are weighted dynamically during the fusion process so that the adversary effect of irrelevant views can be reduced effectively. Here, the similarity in an irrelevant view contributes adversarially to the similarity measurement matching the ground truth. Specifically, in each iteration, there are two steps. In the *semi-supervised metric fusion* step, for each view we form a fused metric by combining the current metric of the view and the label propagation results from other views. Unlike the methods in [2, 4] that obtain a fused metric from multi-views without label information, the metric fusion step in our method fully utilizes the label information from all views. In the *label propagation* step, in each view we conduct label propagation using the fused metric. This step incorporates the complementary information from other views rather than from a single view only. Our SSMF method iteratively conducts the two steps until convergence.

The critical idea here is that the metric fusion and cross-view label propagation processes are complementary to each other. Moreover, we fuse the similarity matrix from one view and label the relevance matrix from other views to yield a cross-view based query (label) driven similarity matrix.

Contributions. Our major contributions can be summarized as follows.

1. We develop an iterative fusion approach SSMF in this paper. SSMF fuses metrics and label propagation results from multi-views iteratively until the fused metric and label propagation results converge simultaneously. We prove the convergence in SSMF theoretically.
2. To further improve the performance of SSMF, we extend it to WSSMF, a novel strategy that automatically generates different weight parameters to views in the fusion process. WSSMF effectively addresses the problem of irrelevant views that are undesirable to fuse in the fusion process for each iteration.
3. Our comprehensive experiments on real image data sets show that our techniques significantly outperform the state-of-the-art methods in terms of accuracy evaluated by varied metrics.

2 Related Work

To our best knowledge, our proposed technique is the first co-training based method for multi-view and graph-based semi-supervised learning problem. Existing one-combo-fits-all methods linearly and independently combine either the metric (kernel) or the labeling propagation result from multiple views to yield a better performance than single view paradigms, as introduced in section 1.

We remark that wang *et al.* [4] proposed a related Unsupervised based Metric Fusion (UMF for short) method. One may adapt UMF for multi-view based semi-supervised learning problem straightforwardly. However, There are several fundamental differences between this straightforward adaption of UMF [4] and SSMF. **First**, it fuses equal weight as suggested by UMF. **Second and foremost**, Unlike the adapted UMF that fuses the pair-wise similarity metric information, which cannot utilize the graph structure to evaluate the similarity between pair-wise objects. SSMF fuses label propagation and similarity metric information interactively for each view and at each iteration, the label propagation can be regarded as a variant of graph random walk, which effectively utilize the graph structure to produce better similarity values among objects than simply fusing the initial pair-wise similarity suggested by UMF. **Finally**, the existence of irrelevant views may significantly affect learning results. While the UMF paradigm is unable to distinguish the different contributions of different views in a fusion based learning process, we devise an effective process to iteratively identify the weights of views by taking the advantage of availability of label propagation result at each time stamp. Wang *et al.* [7] proposed another metric fusion technique against multi-view data via a cross-view based graph random walk approach, however, they studied the unsupervised case rather than semi-supervised learning studied in this paper.

3 SSMF

In this section, we present SSMF and describe its two nice properties, namely ***global consistency and local smoothness*** [8]. We first review the prelimi-

naries. Then, we discuss SSMF using two views. Last, we present the general iterative form of SSMF with multi-views.

3.1 Preliminaries

Let $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ be a set of data points from $\mathbf{M}$ views, we construct $\mathbf{M}$ graphs each using a different feature. $\mathbf{G}^g$ denotes a k-NN graph constructed on $\mathbf{X}$ using g-th feature. Specifically, $\mathbf{G}^g$ is constructed by connecting every two vertices x_i and x_j if one is among the k nearest neighbors of the other. Here, the nearest neighbors are computed using Euclidean distance between the g-th feature vectors of the images. The Euclidean distance between the g-th feature vectors of x_i and x_j is denoted as $||x_i, x_j||_g$. $\mathbf{W}_g$ denotes the edge affinity matrix of $\mathbf{G}^g$. Each entry $\mathbf{W}_g(i,j)$ in $\mathbf{W}_g$ represents the similarity between x_i and x_j according to the g-th feature vector. $\mathbf{W}_g(i,j)$ is defined by a Gaussian kernel and is set to

$$\mathbf{W}_g(i,j) = exp(-||x_i, x_j||_g^2/2\sigma^2) \tag{1}$$

if there is an edge in $\mathbf{G}^g$ between x_i and x_j. Otherwise, $\mathbf{W}_g(i,j)$ is zero. $\mathbf{D}_g$ is the diagonal matrix of $\mathbf{G}^g$ where each element $\mathbf{D}_g(i,i)$ is defined as $\mathbf{D}_g(i,i) = \sum_{j=1}^{n} \mathbf{W}_g(i,j)$.

Without loss of generality, assume the first m points x_i $(i = 1, 2, \ldots, m)$ are labeled points and the remaining points are unlabeled. Let the number of labels be c, and $\mathbf{L} \in \mathbb{R}^{n \times c}$ be the relevance labeling matrix with $\mathbf{L}(i,j) = 1$, if x_i is labeled by label j, denoted by $\mathbf{L}(x_i) = j$ $(1 \leq j \leq c)$, and 0 otherwise. Here, we assume each point is associated with a single class label from the label set. Similarly, let $\mathbf{R}_g \in \mathbb{R}^{n \times c}$ be the relevance score of unlabeled point x_u belonging to label j regarding the g-th view. The closed form of optimal $\mathbf{R}_g$ is yielded by minimizing the objective function

$$\mathbf{F}(\mathbf{R}_g) = \frac{1}{2}\Big(\sum_{i,j=1}^{n} \mathbf{W}_g(i,j)\Big(\frac{1}{\sqrt{\mathbf{D}_g(i,i)}}(\mathbf{R}_g(i,\cdot) \tag{2}$$

$$- \frac{1}{\sqrt{\mathbf{D}_g(j,j)}}(\mathbf{R}_g(j,\cdot))^2 + \alpha_g \sum_{i=1}^{n}(\mathbf{R}_g(i,\cdot) - \mathbf{L}(i,\cdot))^2\Big)$$

where $\mathbf{R}_g(i,\cdot)$ and $\mathbf{L}(i,\cdot)$ are the i-th row of $\mathbf{R}_g$ and $\mathbf{L}$, respectively. The first term in the right hand side of Eq. (2) represents the *local smoothness*, which means that $\mathbf{R}_g(i,\cdot)$ is similar to $\mathbf{R}_g(j,\cdot)$ if x_i and x_j are proximate to each other. The second term in Eq. (2) represents the *global consistency*, which means that the final labeling matrix $\mathbf{R}_g$ should be consistent to the initial labeling matrix $\mathbf{L}$. We minimize $\mathbf{F}(\mathbf{R}_g)$ by setting $\frac{\partial \mathbf{F}(\mathbf{R}_g)}{\partial \mathbf{R}_g} = 0$, and have

$$\mathbf{R}_g^\star = (I - \alpha_g \mathbf{S}_g)^{-1}\mathbf{L} \tag{3}$$

where $\mathbf{S}_g = \mathbf{D}_g^{-\frac{1}{2}}\mathbf{W}_g\mathbf{D}_g^{-\frac{1}{2}}$, $\mathbf{D}_g$ is the diagonal matrix with the i-th diagonal element $\mathbf{D}_g(i,i) = \sum_{j=1}^{n} \mathbf{W}_g(i,j)$, and α_g is a real value such that $0 < \alpha_g < 1$. $\mathbf{R}_g^\star$ can also be regarded as the label propagation result on $\mathbf{G}_g$.

3.2 SSMF for Two Views

Instead of directly computing the similarity metric between any pair-wise points under unsupervised scenario [4], we achieve the similarity, under semi-supervised scenario, by indirectly measuring relevance between each point and all labels, formulated as labeling relevance matrix. As such, one can imagine that if both data points have large relevance regarding all labels, their similarity is large, otherwise, it is small. In order to learn semi-supervised metric regarding two views, we need to consider the following two challenges. That is, (1) the learned similarity metric should encode the relevance between data points and all labels. (2) the learned metric should well incorporate the complementary information from two views to achieve the consistency. Assume $\mathbf{W}_g^{[t+1]}$ $(g = 1, 2)$ denote the metric similarity matrix for g-th view in $t + 1$ iterations, then we define the following semi-supervised fusion strategy:

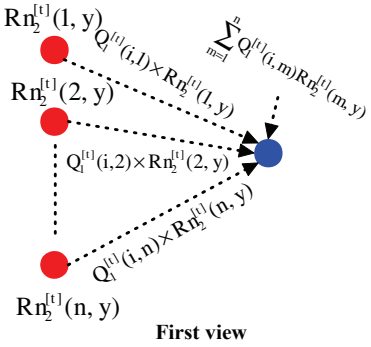


Fig. 1. Semi-supervised metric fusion step regarding the first view formulated as Eq. (4).

$$\mathbf{W}_1^{[t+1]} = \mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]}(\mathbf{Rn}_2^{[t]})^T(\mathbf{Q}_1^{[t]})^T + \lambda I \tag{4}$$

$$\mathbf{W}_2^{[t+1]} = \mathbf{Q}_2^{[t]}\mathbf{Rn}_1^{[t]}(\mathbf{Rn}_1^{[t]})^T(\mathbf{Q}_2^{[t]})^T + \lambda I \tag{5}$$

where $\mathbf{Q}_g^{[t]}$ $(g = 1, 2)$ is the normalized affinity matrices such that $\mathbf{Q}_g^{[t]}(i,j) = \frac{\mathbf{W}_g^{[t]}(i,j)}{\sum_{j=1}^{n}\mathbf{W}_g^{[t]}(i,j)}$, $\mathbf{Rn}_g^{[t]}(i,j) = \frac{\mathbf{R}_g^{[t]}(i,j)}{\sum_{j=1}^{c}\mathbf{R}_g^{[t]}(i,j)}$, the goal of using normalized form is to avoid the huge difference in scale of the label relevance matrices in different views. I is identity matrix, and λI is incorporated to make SSMF robust to the noise. To better explain the above fusion strategy, we take Eq. (4) as an example of refining the metric for the first view by applying SSMF, as illustrated in Fig. 1.

Intuition. We divide the right-hand-side of Eq. (4) into two parts, as $\mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]}$, and its transpose $(\mathbf{Rn}_2^{[t]})^T(\mathbf{Q}_1^{[t]})^T$, we study each entry of $\mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]}$ for any iteration t

$$(\mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]})(i,y) = \sum_{m=1}^{n}\mathbf{Q}_1^{[t]}(i,m)\mathbf{Rn}_2^{[t]}(m,y) \tag{6}$$

As illustrated in Fig. 1, $(\mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]})(i,y)$ represents the fused relevance scores between the y-th label and x_i in the first view, which can be seen as the summation of propagation of label relevance score between x_m and y-th label formulated as $\mathbf{Rn}_2^{[t]}(m,y)$, through the edge weight equivalent to similarity between x_i and x_m $(m \neq i)$, formulated as $\mathbf{Q}_1^{[t]}(i,m)$ to x_i. Such $(\mathbf{Q}_1^{[t]}\mathbf{Rn}_2^{[t]})(i,y)$ is obtained by incorporating the metric, $\mathbf{Q}_1^{[t]}(i,m)$, $m \neq i$, from the first view, and label relevance matrix, $\mathbf{Rn}_2^{[t]}(m,y)$, $m \neq i$, from the second view to make the incorporation of the complementary information from two views. Following this principle, the refined $\mathbf{W}_1^{[t+1]}(i,j)$ in next iteration $t + 1$, for the first view, is yielded by considering relevance score between all labels and both two points

(x_i and x_j, respectively), while effectively incorporates the complementary information from two views. Eq. (5) may be conducted similarly. ■

One natural question is how to calculate $\mathbf{R}_g^{[t]}$, and its normalized form $\mathbf{Rn}_g^{[t]}$ for each iteration t, we propose to adopt the general iterative form in the next section. It has the following two nice features: (1) calculating $\mathbf{R}_g^{\star}$ via Eq. (3) is costly due to inverse matrix computation. (2) iterative label propagation is essentially relying on the graph random walk, which will improve the metric similarity between points and points (and label-data relevance) by exploring the graph structure.

3.3 The General Iterative Form of SSMF

We can get $\mathbf{R}_g^{[t]}$ $(g = 1, 2, \ldots, M)$ iteratively by

$$\mathbf{R}_g^{[t]} = \alpha_g \mathbf{P}_g^{[t]} \mathbf{R}_g^{[t-1]} + (1 - \alpha_g)\mathbf{L} \tag{7}$$

where $\mathbf{P}_g^{[t]}(i,j) = \frac{\mathbf{W}_g^{[t]}(i,j)}{\mathbf{D}_g^{[t]}(i,i)+\mathbf{D}_g^{[t]}(j,j)}$ $(g = 1, 2, \ldots, M)$, and $0 \leq \alpha_g \leq 1$. $\mathbf{P}_g^{[t]}$ is a symmetric matrix. $\mathbf{L}$ is the initial labeling matrix mentioned in Section 3.1.

Generalizing Eqs. (4) and (5) regarding two views, $\mathbf{W}_g^{[t]}$ may be calculated as follows for multi-views.

$$\mathbf{W}_g^{[t+1]} = \mathbf{Q}_g^{[t]} \left(\frac{\sum_{j \neq g} \mathbf{Rn}_j^{[t]}}{M-1}\right)\left(\frac{\sum_{j \neq g} (\mathbf{Rn}_j^{[t]})^T}{M-1}\right)(\mathbf{Q}_g^{[t]})^T + \lambda I \tag{8}$$

The iterative form of SSMF with multi-view by iteratively applying Eqs. (7) and (8) represents a novel label propagation process. Specifically, each weighted graph $\mathbf{G}_g^{[t]}$ associated with the matrix $\mathbf{W}_g^{[t]}$ or $\mathbf{P}_g^{[t]}$, incorporates the label propagation results inherent in $\mathbf{Rn}_j^{[t]}$ $(j \neq g)$ from other views, as shown in Eq. (8), and hence we call the label propagation formulated as Eq. (7) as ***cross view label propagation***.

Now, we are ready to prove the convergence of SSMF.

Theorem 1. *The iterative form of SSMF formulated in Eq.* (7) *converges.*

It suffices to prove the convergence on one view. Following Eq. (7), we have

$$\mathbf{R}_g^{[t]} = \alpha_g^t \mathbf{L} \prod_{i=1}^{t} \mathbf{P}_g^{[i]} + (1 - \alpha_g)\mathbf{L} \sum_{i=1}^{t-1} \alpha_g^i \prod_{j=1}^{i} \mathbf{P}_g^{[j]} \tag{9}$$

where $\mathbf{R}_g^{[0]} = \mathbf{L}$. Apparently, since $0 < \alpha_g < 1$,

$$\lim_{t \to \infty} \alpha_g^t \mathbf{L} [\prod_{i=1}^{t} \mathbf{P}_g^{[i]}](i,j) = 0$$

Moreover, the largest eigenvalue of $\mathbf{P}_g^{[i]}$ $(i = 1, 2, \ldots, t)$ is no more than 1 according to the Gershgorin circle theorem. For the second term in Eq. (9), $(1 - \alpha_g)\mathbf{L}$

Algorithm 1: The algorithm of SSMF

Input: Initial affinity matrix $W_g^{[1]}(g = 1, 2, \cdot\cdot, \cdot, M)$, $R_g^{[0]}$, α_g, λ , initial label relevance matrix L in Eq. (7), the convergence threshold ϵ.

Output: The final label relevance matrix $R^{\star}_{SSMF}$

for $g = 1, \cdots, M$ **do**
  $t = 0$.
  Obtaining the label propagation $R_g^{[1]}$ by Eq. (7)
$t = 1$.
repeat
  for $g = 1, \cdots, M$ **do**
    $\mathbf{Z}_g^{[t]} = \mathbf{Q}_g^{[t]}(\frac{\sum_{j\neq g}\mathbf{R}\mathbf{n}_j^{[t]}}{M-1})$
    $\mathbf{W}_g^{[t+1]} = \mathbf{Z}_g^{[t]}(\mathbf{Z}_g^{[t]})^T + \lambda I$
    $\mathbf{R}_g^{[t+1]} = \alpha_g \mathbf{P}_g^{[t+1]}\mathbf{R}_g^{[t]} + (1-\alpha_g)\mathbf{L}$
    $t = t + 1$;
until *change is smaller than* ϵ;
$R^{\star}_{SSMF} = \sum_{g=1}^{M} \frac{(\mathbf{R}_g)^{\star}_{SSMF}}{M}$;
// $(\mathbf{R}_g)^{\star}_{SSMF}$ is the converged relevance label matrix in the g-th view.

is a constant matrix for all $\alpha_g^i[\prod_{j=1}^{i}\mathbf{P}_g^{[j]}]$ at any step i, thus, we only need to consider the series $\sum_{i=1}^{t-1}\alpha_1^i[\prod_{j=1}^{i}\mathbf{P}_g^{[j]}]$. We denote the i-th term by $\mathcal{H}_g[i]$ and study the convergence of entry $\mathcal{H}_g[i](l,m)$. We only need to prove the convergence of series $\sum_{i=1}^{t-1}\mathcal{H}_g[i](l,m)$, where $\mathcal{H}_g[i](l,m) = \alpha_g^i[\prod_{j=1}^{i}\mathbf{P}_g^{[j]}](l,m) < \alpha_g^i$, since $[\prod_{j=1}^{i}\mathbf{P}_g^{[j]}](l,m) < 1$, which can be easily verified by simple arithmetic operations. We construct the series $\sum_{i=1}^{t-1}\alpha_g^i$ $(0 < \alpha_g < 1)$. Obviously, the series converges, since $[\mathcal{H}_g[i]](l,m) \leq \alpha_g^i$ and each item $[\mathcal{H}_g[i]](l,m)$ is positive. ■

Let $(\mathbf{R}_g)^{\star}_{SSMF}$ be the convergent label relevance matrix regarding the g-th view by interactively applying Eq. (7) (cross-view label propagation) and Eq. (8) (semi-supervised metric fusion). The final label relevance matrix regarding multi-views is $\mathcal{R}^{\star}_{SSMF} = \sum_{g=1}^{M}\frac{(\mathbf{R}_g^{\star})_{SSMF}}{M}$. We summarize the algorithm of SSMF in Algorithm 1.

One important issue that SSMF does not consider is that there may be some *irrelevant views*, and simply fusing all views using the same weight in Eq. (8) may not achieve the best overall performance if there are irrelevant views during the fusion process. To address this issue, we devise an effective learning method to assign a weight to each view in each fusion iteration. Consequently, we extend SSMF to WSSMF, which will be described in next section.

4 WSSMF: Learning Weights for SSMF

The basic idea is to consider the labeling result of cross-view label propagation for unlabeled points in the set $\mathbf{U}$ in each iteration. Two views are regarded

consistent if their labeling results are similar. Specifically, we denote by $\mathbf{V}_i^{[t]}$ the i-th view in iteration t. The more consistent $\mathbf{V}_i^{[t]}$ and $\mathbf{V}_j^{[t]}$ $(1 \leq j \neq i \leq M)$ are, the larger the weight parameter $\theta_{ij}^{[t]}$ is for $\mathbf{V}_j^{[t]}$. Note that the labeling result of cross-view label propagation may be different at various iterations. Therefore, we calculate the weight parameter in different iterations. We define a function $\mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_j^{[t]})$ in Eq. (10) to measure the mismatch between i-th and j-th view in terms of cross-view labeling propagation result.

$$\mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_j^{[t]}) = \sum_{x_u^{[0]} \in \mathbf{U}, \mathrm{L}(x_u^{[t]}) \neq 0} \mathbf{B}(\mathrm{L}(x_u^{[t]}[i]), \mathrm{L}(x_u^{[t]}[j])) \tag{10}$$

where $\mathbf{B}(\mathrm{L}(x_u^{[t]}[i]), \mathrm{L}(x_u^{[t]}[j])) = ||\mathrm{L}(x_u^{[t]}[i]) - \mathrm{L}(x_u^{[t]}[j])||$, $||\cdot||$ is the absolute value operator, and $\mathrm{L}(x_u^{[t]})$ is the largest label relevance score of $x_u^{[t]}$ regarding all labels. We have $\mathrm{L}(x_u^{[t]}[i]) = \max_l\{\mathbf{Rn}_i^{[t]}(u, l)\}$.

Initially, we set the label relevance score of all unlabeled points to be 0, and $\mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_j^{[t]})$ describes the inconsistency degree between $\mathbf{V}_i^{[t]}$ and $\mathbf{V}_j^{[t]}$ at iteration t. The larger $\mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_j^{[t]})$ is, the more inconsistent $\mathbf{V}_i^{[t]}$ and $\mathbf{V}_j^{[t]}$ are. For $\mathbf{V}_i^{[t]}$, the weight parameter $\theta_{ij}^{[t]}$ $(i \neq j)$ for $\mathbf{V}_j^{[t]}$ is defined as

$$\theta_{ij}^{[t]} = 1 - \frac{\mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_j^{[t]})}{\sum_{h \neq i} \mathbf{D}(\mathbf{V}_i^{[t]}, \mathbf{V}_h^{[t]})} \tag{11}$$

Immediately, we have $\theta_{ij}^{[t]} = \theta_{ji}^{[t]}$ and $0 \leq \theta_{ij}^{[t]} \leq 1$. They are the entries in the coefficient symmetric matrix in iteration t, denoted by $\Theta^{[t]}$. In iteration t, the j-th view $(1 \leq j \neq i \leq M)$ is said to be ***irrelevant*** with respect to the i-th view if $\theta_{ij}^{[t]} < \frac{\sum_{g \neq i} \theta_{ig}^{[t]}}{M-1}$, otherwise, the j-th view is said to be ***relevant***. For the i-th view, we denote ***the set of relevant views*** at iteration t by $\mathbf{Re}_i^{[t]}$.

Instead of computing global irrelevant views explicitly, for the i-th view, we only fuse the views from $\mathbf{Re}_i^{[t]}$ in iteration t, and set the correlation strength weight to be 0 for irrelevant views. Combining Eq. 11 and Eq. 8, we have the Weighted SSMF (WSSMF for short) for multi-views, which iteratively applies Eq. 7 and Eq. 12 until convergence.

$$\begin{aligned} \mathbf{W}_g^{[t+1]} = \mathbf{Q}_g^{[t]} \Big(& \frac{\sum_{j \in \mathbf{Re}_g^{[t]}} \theta_{gj}^{[t]} \mathbf{Rn}_j^{[t]}}{|\mathbf{Re}_g^{[t]}|} \Big) \times \\ & \Big(\frac{\sum_{j \in \mathbf{Re}_g^{[t]}} (\theta_{gj}^{[t]} \mathbf{Rn}_j^{[t]})^T}{|\mathbf{Re}_g^{[t]}|} \Big) (\mathbf{Q}_g^{[t]})^T + \lambda I \end{aligned} \tag{12}$$

Like SSMF, WSSMF also converges, which can be immediately proved in the same manner as Theorem 1. Therefore, the final optimal label relevance matrix can be obtained as $\mathcal{R}_{WSSMF}^{\star} = \sum_{g=1}^{M} \frac{(\mathbf{R}_g)_{WSSMF}^{\star}}{M}$, where M is the number of views, $(\mathbf{R}_g)_{WSSMF}^{\star}$ is the convergent label relevance matrix in the g-th view

obtained using WSSMF. Based on Algorithm 1, we generate the algorithm of WSSMF by replacing $\mathbf{Z}_g^{[t]}$ in line **7** with $\mathbf{Z}_g^{[t]} = \mathbf{Q}_g^{[t]}(\frac{\sum_{j\in\mathbf{Re}_g^{[t]}} \theta_{gj}^{[t]}\mathbf{Rn}_j^{[t]}}{|\mathbf{Re}_g^{[t]}|})$, and $\mathbf{W}_g^{[t+1]}$ in line **8** with Eq. 12.

4.1 Complexity Analysis

Now, we analyze the time complexity of each iteration in SSMF and WSSMF.

The cost of SSMF mainly comes from two parts: cross-view label propagation and semi-supervised metric fusion. The iterative cross-view label propagation in Line 9 of Algorithm 1 takes $\mathcal{O}(Mn^2c)$ time, and the same time complexity holds for semi-supervised metric fusion in Lines 7-8. We remark that all the above cost is from the matrix multiplication rather than matrix inverse computation. It is well known that matrix multiplication implementation without inverse computation is efficient. Similar to SSMF, WSSMF also needs $\mathcal{O}(Mn^2c)$ time for both metric fusion and cross-view label propagation. In addition, $\mathcal{O}(M^2n)$ time is needed to obtain the view correlation matrix Θ in each iteration regarding M views. Therefore, the overall time complexity for WSSMF is $\mathcal{O}(Mn^2c)+\mathcal{O}(M^2n)$. As observed in our experiments(refer to Fig 3), both SSMF and WSSMF converge within quite limited iterations for most cases (less than 65 times).

5 Experiments

We evaluate both SSMF and WSSMF using multi-view content based image retrieval (CBIR) and multi-label image classification on real data sets. We set the convergence threshold ϵ to 10^{-4} for all methods.

In our experiments, we compare with the following state-of-the-art multi-view graph based methods for both multi-view CBIR and multi-label image classification.

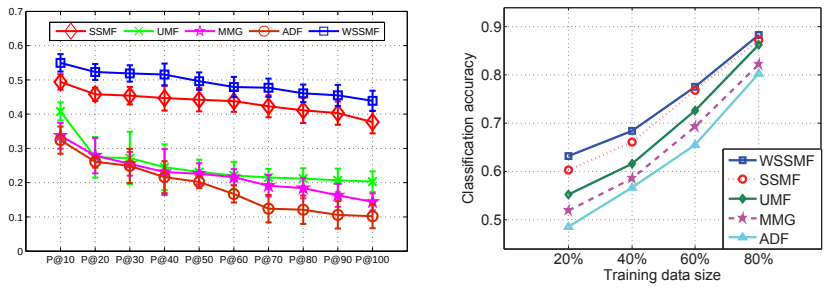
- *The multi-modality graph (MMG) method* [3], which uses multiple graph models under different views. The final ranking score vector is obtained by combining the independent label propagation (manifold ranking) results carried by each image in each view with different weights.
- *The averaged distance of multiple feature based metric (ADF) method* [2], which constructs a single relevance graph using the metric of average distance from multiple views.
- *The unsupervised metric fusion (UMF) method* [4], which conducts metric fusion without considering label propagation result. It is adapted to tackle multi-view graph-based semi-supervised learning as follows. We first obtain the convergent affinity matrix $\mathbf{W}_g$ $(g = 1, 2, \ldots, M)$ for the g-th view by applying UMF, and then obtain the ranking score vector by optimizing Eq. (2), where the affinity matrix $\mathbf{W}_g$ is the fused affinity matrix using UMF on multi-views.

5.1 Multi-view Content Based Image Retrieval (CBIR)

Multi-view CBIR is a typical problem where graph based multi-view semi-supervised learning is extensively applied. Specifically, a query image is a labeled

data object in our model, and the label relevance matrix $R_g \in \mathbb{R}^{n \times c}$ in Eq. (2) is reduced to a ranking score vector $r_g \in \mathbb{R}^n$, and $R_g(i, \cdot) \in \mathbb{R}^n$ is reduced to $r_g(i) \in \mathbb{R}$, which represents the relevance score between x_i and the query image (labeled image). $\mathbf{L} \in \mathbb{R}^{n \times c}$ in Eq. (2) is reduced to an n dimensional vector $\mathbf{Y} \in \mathbb{R}^n$ with the i-th entry to be 1 if x_i is the query image, and 0 otherwise.

We set the number of nearest neighbors k to 20 to calculate the metric distance in Eq. (1) for all views, which is consistent with the UMF method [4]. Similar to [9], we set α_g to 0.99 in Eq. (7) for all views, set λ to 1 in both Eq. (8) and Eq. (12). All methods are tested on the COREL5K data set [1], which consists of 5000 images in 50 categories. Each category contains 100 images. Due to the same number of images in each category, we use the *precision-scope* [3] as the evaluation metric. We use HOG, color histogram, RGB-SIFT and Pyramid wavelet texture feature to construct different views, most of them are utilized by MMG. For each method, we select every sample of 5000 images as the query image (labeled objects) each time, and obtain the average precision value and its statistical distribution regarding all 5000 samples, shown using 3 points (mean, +1 standard deviation, and -1 standard deviation) in Fig. 2(a).



(a) (b)

Fig. 2. (a) Top-s precision on COREL5K data set. (b) Classification accuracy with respect to sample rate on Caltech-101 image data set.

Unsurprisingly, WSSMF outperforms the other methods in top-s average precision, since it can better achieve the consistency from multi-views than the other methods. In addition, it can effectively address the problem of irrelevant views at each iteration. SSMF is the next after WSSMF. SSMF does not handle the problem of irrelevant views. To demonstrate the difference, we take one query image from the "boat" category, and list the coefficient matrix at the seventh iteration, denoted by $\Theta^{[7]}$ obtained by WSSMF, as follows.

$$\Theta^{[7]} = \begin{array}{c|cccc} & \text{HOG} & \text{Texture} & \text{Color} & \text{RGB-SIFT} \\ \hline \text{HOG} & 1 & 0.462 & 0.246 & 0.292 \\ \text{Texture} & 0.462 & 1 & 0.292 & 0.246 \\ \text{Color} & 0.246 & 0.292 & 1 & 0.462 \\ \text{RGB-SIFT} & 0.292 & 0.246 & 0.462 & 1 \end{array}$$

The matrix $\Theta^{[7]}$ indicates that Color histogram and RGB-SIFT are the irrelevant views for HOG. Unlike the other methods, WSSMF sets the weight of irrelevant views to 0 for HOG at this iteration. Like SSMF, UMF **(1)** does not

consider the irrelevant view detection, either. Moreover, **(2)** UMF does not fuse label propagation results during the fusion process, **(3)** as such UMF fails to further exploring the graph structure to improve the metric similarity like SSMF and WSSMF as discussed in section 2. Consequently, UMF is inferior to SSMF.

Both MMG and ADF perform worse than the others. MMG outperforms ADF in most cases, since MMG fully explores the graph structure for different views, and it linearly combines the independent label propagation results with different weights. ADF, however, is different from MMG. It assigns the same weight to all views in combining the label propagation results, the single graph associated with averaged metric obstructs the graph structure of original inherent individual views. However, MMG is inferior to SSMF and WSSMF, since such one-combo-fits-all late fusion method is undesirable to achieve the consistency among all views by independently fusing all the label propagation result from all views. Worse still, it cannot well handle the irrelevant views issue.

Fig. 3 shows the 5-point box-plots (maximum, minimum, mean, +1 standard deviation, and -1 standard deviation) of number of iterations and running time of all queries in all methods. Both WSSMF and SSMF use more iterations on average and sot longer running time than ADF and UMF, because ADF and UMF construct only one similarity graph. Instead, WSSMF, SSMF and MMG construct multiple graphs. WSSMF and SSMF need less iterations on average to reach convergence than MMG, since the cross-view based fusion method can speed up the process of achieving consistency. However, the running time of WSSMF and SSMf is similar to that of MMG, since more matrix multiplication is performed during each iteration than MMG.

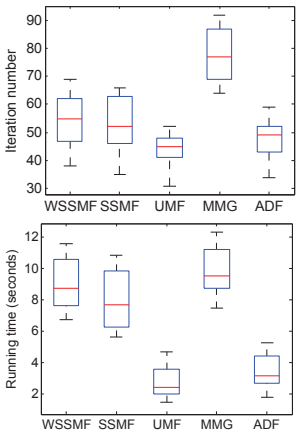


Fig. 3. Comparison of number of iterations (top figure) and running time (bottom figure).

5.2 Multi-view based multi-label image classification

Multi-view based multi-label image classification can be regarded as multi-view based semi-supervised learning with graphs. The Caltech-101 data set (`http://www.vision.caltech.edu/Image_Datasets/Caltech101/`) is used to test multi-label image classification. It contains 9146 images organized into 101 categories. The number of images in different categories ranges from 40 to 800. We set $c = 101$ and $n = 9146$ in the label relevance matrix $R_g \in \mathbb{R}^{n \times c}$ and $\mathbf{L} \in \mathbb{R}^{n \times c}$ in Eq. (2), along with $k = 20$ in Eq. (1) and $\lambda = 1$ in Eq. (12).

We use the same sample rate to draw a random sample of images from each category as labeled images. The rest of images are treated as unlabeled. Each experiment is repeated 5 times, and the average value is reported. The classification accuracy on all unlabeled images is used to evaluate different methods. HOG, color histogram, pyramid wavelet texture feature and SIFT are used to construct different views. The results are shown in Fig. 2(b).

WSSMF outperforms the other methods. SSMF is the second best method. The results verify the advantages of our iterative fusion methods. We also observe that the difference among different methods decreases as the sample rate increases, since a higher sample rate makes the problem less challenging.

6 Conclusion

In this paper, we propose a novel iterative fusion technique for graph based semi-supervised learning from multi-views. The central idea is to fuse metrics and label propagation results from multi-views iteratively and weight views dynamically. The experimental results clearly show that our new methods outperform the state-of-the-art methods on real data sets. As future work, we will investigate how to fuse selective labeling results from multi-view based graphs rather than tackling all the data points including both informative and noise data points. We will also investigate active learning based methods for better effectiveness and efficiency.

Acknowledgment. Jian Pei's Research is supported in part by an NSERC Discovery Grant and a BCFRST NRAS Endowment Research Team Program Project. Xuemin Lin is supported by ARC DP0987557, ARC DP110102937, ARC DP120104168 and NSFC61021004. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. P. Duygulu, K. Barnard, J. D. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.
2. Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *CVPR*, 2010.
3. H. Tong, J. He, M. Li, C. Zhang, and W. Ma. Graph based multi-modality learning. In *ACM MM*, 2005.
4. B. Wang, J. Jiang, W. Wang, Z. Zhou, and Z. Tu. Unsupervised metric fusion by cross diffusion. In *CVPR*, 2012.
5. M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, and Y. Song. Unified video annotation via multigraph learning. *IEEE Trans. Circuits Syst. Video Techn*, 19(5):733–746, 2009.
6. Y. Wang, M. A. Cheema, X. Lin, and Q. Zhang. Multi-manifold ranking: Using multiple features for better image retrieval. In *PAKDD*, 2013.
7. Y. Wang, X. Lin, and Q. Zhang. Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In *ACM CIKM*, 2013.
8. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schlkopf. Learning with local and global consistency. In *NIPS*, 2003.
9. D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schlkopf. Ranking on data manifolds. In *NIPS*, 2003.
10. X. Zhu. Semi-supervised learning with graphs. *PhD thesis,Carnegie Mellon University*, 2005.