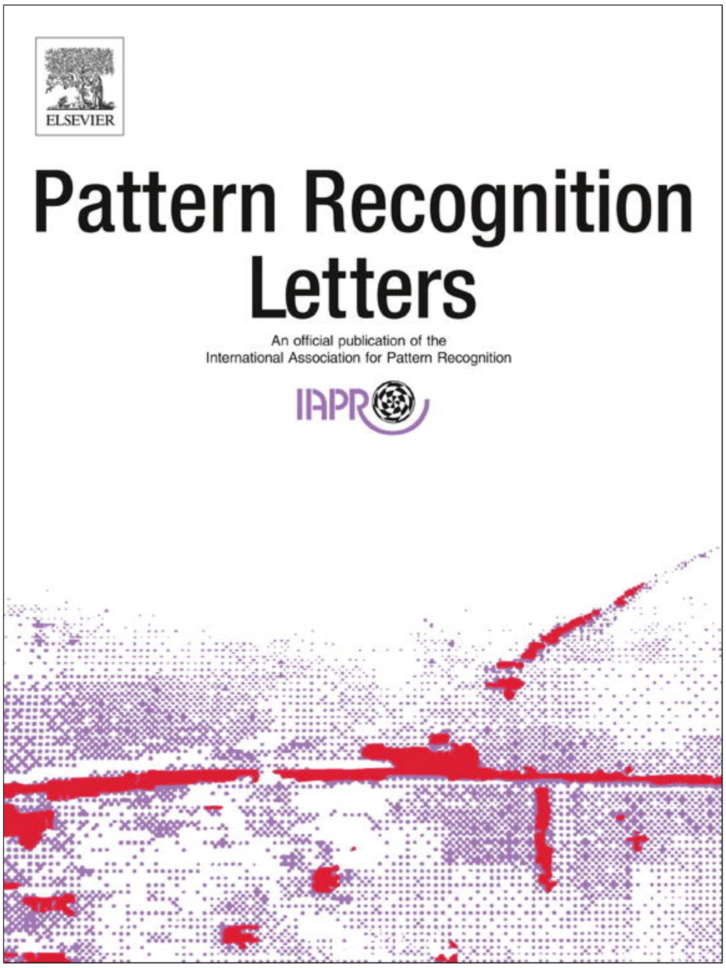

ELSEVIER
Pattern Recognition Letters
An official publication of the
International Association for Pattern Recognition
IAPR

Pattern Recognition Letters 34 (2013) 1263–1269

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Efficiency investigation of manifold matching for text document classification ☆

Ming Sun [a], Carey E. Priebe [b,*]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[b] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA

ARTICLE INFO

Article history:
Received 8 May 2012
1 April 2013
Available online 4 April 2013
Communicated by Y. Liu

Keywords:
Manifold matching
MDS ∘ Procrustes
CCA
JOFC
Efficiency
Classification

ABSTRACT

Manifold matching works to identify embeddings of multiple disparate data spaces into the same low-dimensional space, where joint inference can be pursued. It is an enabling methodology for fusion and inference from multiple and massive disparate data sources. In this paper three methods of manifold matching are considered: PoM, which stands for Multidimensional Scaling (MDS) composed with Procrustes; CCA (Canonical Correlation Analysis) and JOFC (Joint Optimization of Fidelity and Commensurability). We present a comparative efficiency investigation of the three methods for a particular text document classification application.

1. Introduction

1.1. Purpose

In the real world, one single object may have different representations in different domains. For example, the Declaration of Independence has versions translated into different languages. Let n denote the number of objects $O_i, i = 1, \ldots, n$, and K be the number of domains. Then we have

$$\mathbf{x}_{i1} \sim \cdots \sim \mathbf{x}_{ik} \sim \cdots \sim \mathbf{x}_{iK}, \quad i = 1, \ldots, n \tag{1}$$

where the ith object O_i has K measurements $\mathbf{x}_{ik}, k = 1, \ldots, K$; $\mathbf{x}_{ik} \in \Xi_k$ is the representation for object O_i in space Ξ_k.

The problem explored in this paper is that for m new objects $O'_i, i = 1, \ldots, m$, how to classify their representations $\mathbf{y}_{ik} \in \Xi_k$ given the representations $\mathbf{y}_{i'k'} \in \Xi_{k'}$ with $i' \neq i$. For this task, $\mathbf{x}_{ik}, \mathbf{x}_{ik'}, i = 1, \ldots, n$, described above are needed to learn the relation between Ξ_k and $\Xi_{k'}$ so that we can map data from Ξ_k and $\Xi_{k'}$ to a common space χ. Thus $\mathbf{x}_{ik}, \mathbf{x}_{ik'}$ are the domain relation learning training data. This idea is shown in Fig. 1. The domain relation learning training data $\mathbf{x}_{ik}, \mathbf{x}_{ik'}$ are labeled by the filled circles in Ξ_k and $\Xi_{k'}$ respectively. The filled squares in $\Xi_{k'}$ represent the classifier training data $\mathbf{y}_{i'k'}$, which are used to train a classifier g. The classification testing data $\mathbf{y}_{ik}$ is shown by the unfilled square in space Ξ_k. We consider three different domain relation learning methods: PoM, which stands for Multidimensional Scaling (MDS) composed with Procrustes; CCA (Canonical Correlation Analysis) and JOFC (Joint Optimization of Fidelity and Commensurability). We investigate classification performance in the common space χ obtained via PoM, CCA and JOFC, training the classifier on $\mathbf{y}_{i'k'}$ and testing on $\mathbf{y}_{ik}$. The focus of this paper is not on optimizing the classifier; rather, we investigate performance for given classifiers (5-Nearest Neighbor, SVM with degree 2 polynomial kernel) as a function of the number of domain relation learning training data observation n used to learn χ.

1.2. Summary

The structure of the paper is as follows: Section 2 talks about related work. Section 3 discusses the methods employed, including the manifold matching framework as well as embedding and classification details. Experimental setup and results are presented in Section 4. Section 5 is the conclusion.

2. Background

Different methods of transfer learning, multitask learning and domain adaptation are discussed in a recent survey (Pan and Yang, 2010). There are algorithms developed on unsupervised

☆ The Introduction and Background sections and parts of the methodology and experiment setup are taken from Sun et al., 2013 by the same authors. In that paper, Canonical Correlation Analysis and Regularized Canonical Correlation Analysis are compared and contrasted for the same classification task. This paper investigates different manifold matching techniques.

* Corresponding author. Tel.: +1 410 516 7200; fax: +1 410 516 7459.
E-mail addresses: msun8@jhu.edu (M. Sun), cep@jhu.edu (C.E. Priebe).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.03.025

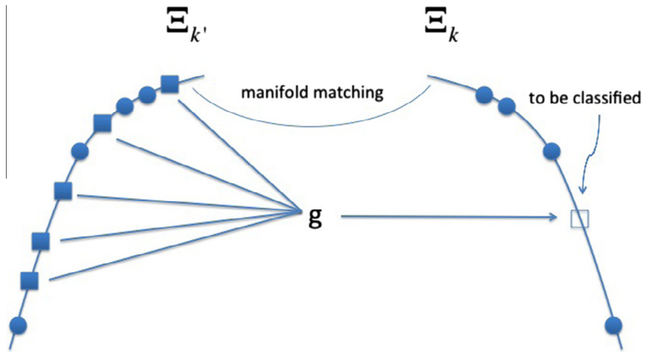


Fig. 1. Classification problem.

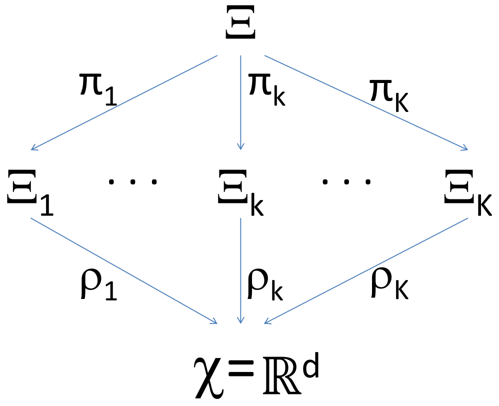


Fig. 2. Manifold matching model.

document clustering where training and testing data are of different kinds (Karakos et al., 2007). The problem explored in this paper can be viewed as a domain adaptation problem, for which the training and testing data of the classifier are from different domains. When the classification is on the text documents in different languages, as described in the later sections of this paper, it is called cross-language text classification. There is much work on inducing correspondences between different language pairs, including using bilingual dictionaries (Olsson et al., 2005), latent semantic analysis (LSA) features (Dumais et al., 1997), kernel canonical correlation analysis (KCCA) (Li and Taylor, 2007), etc. Machine translation is also involved in the cross-language text classification, which translates the documents into a single domain (Rigutini et al., 2005, Fortuna and Shawe-Taylor, 2005, Ling et al., 2008).

The methods investigated in this paper are closely related to various manifold learning and alignment techniques. There has been intensive work done by many people. Principal Components Analysis (PCA) (Jolliffe, 2002) and Multidimensional Scaling (MDS) (Torgerson, 1952; Cox and Cox, 2001; Borg and Groenen, 2005) are classical linear methods to learn low-dimensional representations for high-dimensional observations. In recent years different non-linear manifold learning methods are developed, including kernel PCA (Mika et al., 1999), Isomap (Tenenbaum et al., 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2003), etc. Regarding manifold alignment, Wang and Mahadevan, 2008 applies procrustes for manifold alignment. The use of diffusion maps is discussed in (Lafon et al., 2006). Diaz and Metzler, 2007 learns a common manifold for documents from multilingual corpora such that the embeddings of documents are clustered based on topics. Manifold learning and alignment are also widely used for image analysis (Ham et al., 2006; Wang and Chen, 2009). Lu et al., 2011 presents a discriminative multi-manifold analysis method to solve the single sample per person problem in face recognition. Manifold alignment can be done in a semisupervised way (Ham et al., 2005; Verbeek and Vlassis, 2006), or without pairwise correspondence information (Wang and Mahadevan, 2009; Xiong et al., 2007). In (Pei et al., 2012), unsupervised manifold alignment is conducted based on parameterized distance curves.

3. Method

In this paper, we focus on manifold matching. The whole procedure can be divided into the following steps:

- For each single space Ξ_k, calculate the dissimilarity matrix for all domain relation learning training data observations O_i.
- Run different manifold matching methods on the dissimilarity matrix to get embedding in a common space χ.
- Pursue joint inference (i.e. classification) in the common space χ.

3.1. Manifold matching framework

The framework structure for manifold matching is shown in Fig. 2 (Ma et al., 2012; Priebe et al., in press). For each of the n objects $O_i \in \Xi, i = 1, \ldots, n$, there are K representations $\mathbf{x}_{ik} \in \Xi_k, k = 1, \ldots, K$ generated by the mappings π_k. Manifold matching works to find $\rho_1, \ldots, \rho_K$ to map $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iK}$ to a low-dimensional common space $\chi = \mathbb{R}^d$:

$$\tilde{\mathbf{x}}_{ik} = \rho_k(\mathbf{x}_{ik}), i = 1, \ldots, n, k = 1, \ldots, K. \tag{2}$$

After learning the ρ_ks, we can map a new measurement $\mathbf{y}_k \in \Xi_k$ into the common space $\chi = \mathbb{R}^d$ via:

$$\tilde{\mathbf{y}}_k = \rho_k(\mathbf{y}_k) \tag{3}$$

This allows joint inference to proceed in $\mathbb{R}^d$.

3.2. Embedding

The work described in this paper is based on dissimilarity measures. Let δ_k denote the dissimilarity measure in the kth space Ξ_k, and $\tilde{\delta}$ be the Euclidean distance in the common space $\mathbb{R}^d$. There are two kinds of mapping errors induced by the ρ_ks: fidelity error and commensurability error.

Fidelity measures how well the original dissimilarities is preserved in the mapping $\mathbf{x}_{ik} \mapsto \tilde{\mathbf{x}}_{ik}$, and the fidelity error is defined as the within-condition squared error:

$$\epsilon^2_{f_k} = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} (\tilde{\delta}(\tilde{\mathbf{x}}_{ik}, \tilde{\mathbf{x}}_{jk}) - \delta_k(\mathbf{x}_{ik}, \mathbf{x}_{jk}))^2 \tag{4}$$

Commensurability measures how well the matchedness is preserved in the mapping, and the commensurability error is defined as the between-condition squared error:

$$\epsilon^2_{c_{k_1 k_2}} = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (\tilde{\delta}(\tilde{\mathbf{x}}_{ik_1}, \tilde{\mathbf{x}}_{ik_2}))^2 \tag{5}$$

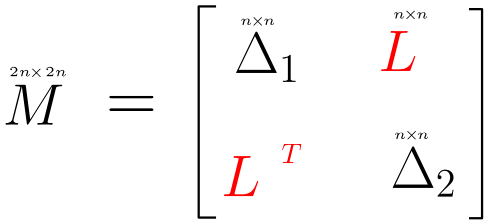

$\overset{2n\times 2n}{M} = \begin{bmatrix} \overset{n\times n}{\Delta_1} & \overset{n\times n}{L} \\ L^T & \overset{n\times n}{\Delta_2} \end{bmatrix}$

Fig. 3. Omnibus dissimilarity matrix.

3.2.1. Procrustes ∘ MDS (P∘M)

Multidimensional Scaling (MDS) (Torgerson, 1952; Cox and Cox, 2001; Borg and Groenen, 2005) works to get a Euclidean representation while approximately preserving the dissimilarities. Given the $n \times n$ dissimilarity matrix $\Delta_k = [\delta_k(\mathbf{x}_{ik}, \mathbf{x}_{jk})]$ in space Ξ_k, multidimensional scaling generates embeddings $\tilde{\mathbf{x}}'_{ik} \in \mathbb{R}^{d'}$ for $\mathbf{x}_{ik} \in \Xi_k, i = 1, \ldots, n, k = 1, \ldots, K$, which attempts to optimize fidelity.

For the $K = 2$ case, multidimensional scaling generates $n \times d'$ matrices $\widetilde{X}'_1$ from Δ_1 and $\widetilde{X}'_2$ from Δ_2. The ith row vector $\tilde{\mathbf{x}}'_{ik}$ of $\widetilde{X}'_k$ is the multidimensional scaling embedding for $\mathbf{x}_{ik}$.

Procrustes works to get the mapping matrix Q^* which satisfies

$$Q^* = \arg\min_{Q^TQ=I} \|\widetilde{X}'_1 - \widetilde{X}'_2 Q\|_F. \tag{6}$$

For the new data $\mathbf{y}_k, k = 1, 2$, based on $\delta_k(\mathbf{y}_k, \mathbf{x}_{ik}), i = 1, \ldots, n$, out-of-sample embedding (Anderson and Robinson, 2003; Trosset and Priebe, 2008) produces $\mathbf{y}_k \mapsto \tilde{\mathbf{y}}'_k$ with $d(\tilde{\mathbf{y}}'_k, \tilde{\mathbf{x}}'_{ik})$ being close to $\delta_k(\mathbf{y}_k, \mathbf{x}_{ik})$. The final embeddings for $\mathbf{y}_1$ and $\mathbf{y}_2$ in the common space $\mathbb{R}^d$ are given by $\tilde{\mathbf{y}}_1 = \tilde{\mathbf{y}}'_1$ and $\tilde{\mathbf{y}}_2 = ((\tilde{\mathbf{y}}'_2)^T Q^*)^T$.

P∘M optimizes fidelity without regard for commensurability (Priebe et al., in press).

3.2.2. Canonical Correlation Analysis (CCA)

Canonical correlation analysis (Hardoon et al., 2004; Hotelling, 1936; Kettenring, 1971) is applied to the multidimensional scaling results. Canonical correlation works to find $d' \times d$ matrices $U_1 : \widetilde{X}'_1 \mapsto \widetilde{X}_1$ and $U_2 : \widetilde{X}'_2 \mapsto \widetilde{X}_2$ as the linear mapping method to maximize correlation for the mappings into $\mathbb{R}^d$.

For new data $\mathbf{y}_k, k = 1, 2$, out-of-sample embedding for multidimensional scaling generates d' dimensional column vector $\tilde{\mathbf{y}}'_k$. The final embeddings in the common space $\mathbb{R}^d$ are given by $\tilde{\mathbf{y}}_1 = U_1^T \tilde{\mathbf{y}}'_1$ and $\tilde{\mathbf{y}}_2 = U_2^T \tilde{\mathbf{y}}'_2$.

Canonical correlation analysis optimizes commensurability without regard for fidelity (Priebe et al., in press). For our work, first we use multidimensional scaling to generate a fidelity-inspired Euclidean representation, and then we use canonical correlation analysis to enforce low dimensional commensurability.

3.2.3. Omnibus Embedding (JOFC)

The omnibus embedding method described in (Priebe et al., in press,Ma et al., 2012) jointly optimizes fidelity and commensurability. Given the $n \times n$ dissimilarity matrices $\Delta_1 \in \Xi_1$ and $\Delta_2 \in \Xi_2$, the $2n \times 2n$ omnibus dissimilarity matrix M is constructed as shown in Fig. 3.

The off-diagonal block is $L = (\Delta_1 + \Delta_2)/2$. The embeddings $\tilde{\mathbf{x}}_{i1}, \ \tilde{\mathbf{x}}_{i2} \in \mathbb{R}^d$ can be obtained by running multidimensional scaling on M directly.

Similar to P∘M and CCA, for the new data $\mathbf{y}_k, k = 1, 2$, its embedding in the common space $\tilde{\mathbf{y}}_k \in \mathbb{R}^d$ can be obtained directly from out-of-sample embedding based on $\delta_k(\mathbf{y}_k, \mathbf{x}_{ik}), i = 1, \ldots, n$.

3.3. Classification

Given the measurements of m new data points $\mathbf{y}_{ik} \in \Xi_k, i = 1, \ldots, m$, for the classification of $\mathbf{y}_{ik}$, we consider the problem in which there are no training data available in Ξ_k and we must borrow training data from another space $\Xi_{k'}$. Let $\mathbf{y}_{i'k'} \in \Xi_{k'}$ denote the training data. The classification procedure begins with projecting both testing data $\mathbf{y}_{ik}$ and training data $\mathbf{y}_{i'k'}$ to a common space $\mathbb{R}^d$. The manifold matching methods PoM, CCA and JOFC described in Section 3.2 embed $\mathbf{y}_{ik} \mapsto \tilde{\mathbf{y}}_{ik} \in \mathbb{R}^d$ and $\mathbf{y}_{i'k'} \mapsto \tilde{\mathbf{y}}_{i'k'} \in \mathbb{R}^d$. As a result, a classifier is trained on $\tilde{\mathbf{y}}_{i'k'}$ and tested on $\tilde{\mathbf{y}}_{ik}$. This problem is motivated by the fact that in many situations there is a lack of training data in the space where the testing data lie. We will discuss the classification problem in more details in Section 4.4.

3.4. Efficiency investigation

We investigate the effect of the number of domain relation learning training data observations on the classification performance. That is, given a subset of available domain relation learning training data, we are interested in how different manifold matching techniques perform in the cross-domain classification task. The classification accuracy is expected to improve with increasing amount of domain relation learning training data. To achieve the same classification accuracy, the method using the smallest amount of domain relation learning training data is identified as the most efficient one.

4. Experiments

In this section the experimental details are described. Section 4.1 describes the dataset used for our experiments. Section 4.2 discusses the dissimilarity matrix calculation. Our method for choosing proper embedding dimensions is presented in Section 4.3. The classification setting and results are described and analyzed in Section 4.4 and Section 4.5.

4.1. Dataset

Our experiments apply different manifold matching techniques (PoM, CCA and JOFC) to text document classification. The dataset is obtained from wikipedia, an open-source multilingual web-based encyclopedia with around 26 million articles in more than 280 languages. Each document may have links pointing to other documents in the same language which explain certain terms in its content as well as the documents in other languages for the same subject. Articles of the same subject in different languages are not necessarily the exact translations of one another. They can be written by different people and their contents can differ significantly.

English articles within a 2-neighborhood of the English article "Algebraic Geometry" are collected. The corresponding French documents of those English ones are also collected. So this data set can be viewed as a two space case: Ξ_1 is the English space and Ξ_2 is the French space. There are in total 1382 documents in each space. That is, $\mathbf{z}_{1,1}, \ldots, \mathbf{z}_{1382,1} \in \Xi_1$, and $\mathbf{z}_{1,2}, \ldots, \mathbf{z}_{1382,2} \in \Xi_2$. Note that $\mathbf{z}_{ik}, i = 1, \ldots, 1382, k = 1, 2$ includes both domain relation learning training data $\mathbf{x}_{ik}, i = 1, \ldots, n$ and new data points $\mathbf{y}_{ik}, i = 1, \ldots, m$ $(m + n = 1382)$ used for classification training and testing.

All 1382 documents are manually labeled into five disjoint classes (0–4) based on their topics. Topics are category, people, locations, date and math respectively. Documents in classes 0, 2, 4 are the domain relation learning training data $\mathbf{x}_{ik}, i = 1, \ldots, n, k = 1, 2$. There are in total 819 documents in those three classes ($n = 819$). The rest 563 ($m = 563$) documents in classes 1, 3 are the new data $\mathbf{y}_{ik}, i = 1, \dots, m, k = 1, 2$. They are used to train a classifier and run the classification test.

4.2. Dissimilarity matrix

The method described in Section 3.2 starts with the dissimilarity matrix. For our work two different kinds of dissimilarity measures are considered: text content dissimilarity matrix Δ_k^t and graph topology dissimilarity matrix Δ_k^g. Both matrices are of dimension 1382×1382, containing the dissimilarity information for all data points $\mathbf{z}_{1k}, \dots, \mathbf{z}_{1382k}$.

Graphs $G_k(V, E_k)$ can be derived from the dataset; V represents the set of vertices which are the 1382 wikipedia documents, and E_k is the set of edges connecting those documents in language k.

The (i,j) entry $\Delta_k^g(i,j) \in \Delta_k^g$ is the number of steps on the shortest path from document i to document j in G_k. In the English space $\Xi_1, \Delta_1^g(i,j) \in \{0, \dots, 4\}$, where the upperbound value 4 comes from the 2-neighborhood document collection. In the French space $\Xi_2, \mathbf{z}_{i2}$ is the French corresponding document for the English one $\mathbf{z}_{i1} \in \Xi_1$, and $\Delta_2^g(i,j) \in \Delta_2^g$ depends on the French graph connections. It is possible that $\Delta_2^g(i,j) \neq \Delta_1^g(i,j)$. At the extreme end, $\Delta_2^g(i,j) = \infty$ when $\mathbf{z}_{i2}$ and $\mathbf{z}_{j2}$ are not connected. We set $\Delta_2^g(i,j) = 6$ for $\Delta_2^g(i,j) > 4$. Because the upperbound of the English graph topology dissimilarity is 4, this choice makes the French graph topology dissimilarity comparable to the English one. On one hand, the French graph topology dissimilarity upperbound is larger since the actual French graph topology dissimilarity can be larger than the English one. On the other hand, the French graph topology dissimilarity upperbound is set to be a value not too big to make sure it does not overwhelm the embedding for the French graph topology dissimilarity matrix. Optimal pre-processing to put the dissimilarities on the same footing is the subject of ongoing investigation.

Table 1
MDS dimensions.

n'	% Of n	d'
82	10	40
164	20	80
246	30	100
328	40	100
410	50	150
491	60	150
573	70	150
655	80	200
737	90	200
819	100	200

$\Delta_k^t(i,j) \in \Delta_k^t$ is based on the text processing features for documents $\mathbf{z}_{ik}, \mathbf{z}_{jk} \in \Xi_k$. Given the feature vectors $\mathbf{f}_{ik}, \mathbf{f}_{jk}, \Delta_k^t(i,j)$ is calculated by the cosine dissimilarity $\Delta_k^t(i,j) = 1 - \frac{\mathbf{f}_{ik} \cdot \mathbf{f}_{jk}}{\|\mathbf{f}_{ik}\|_2 \|\mathbf{f}_{jk}\|_2}$. For our experiments, we consider term frequency–inverse document frequency (TFIDF) features (Salton and Buckley, 1988) as $\mathbf{f}$.

We use multidimensional scaling to embed into Euclidean space $\mathbb{R}^d$ while approximating dissimilarity information; in this space, Euclidean distance is appropriate.

4.3. Embedding dimension selection

To choose the dimension d for the common space $\mathbb{R}^d$, we pick a sufficiently large dimension and embed Δ_k^t and Δ_k^g via multidimensional scaling. The sqare root information of the embedding's covariance matrix is shown in Fig. 4.

Based on the plots in Fig. 4, we choose the dimension $d = 15$ (dimension of the joint space $\chi = \mathbb{R}^d$), which is low but preserves most of the variance (Jolliffe, 2002). This model selection choice of dimension is an important issue in its own right; for this paper, we fix $d = 15$ throughout. The focus of this paper is not model selection. Our selection of d generates satisfactory experimental re-

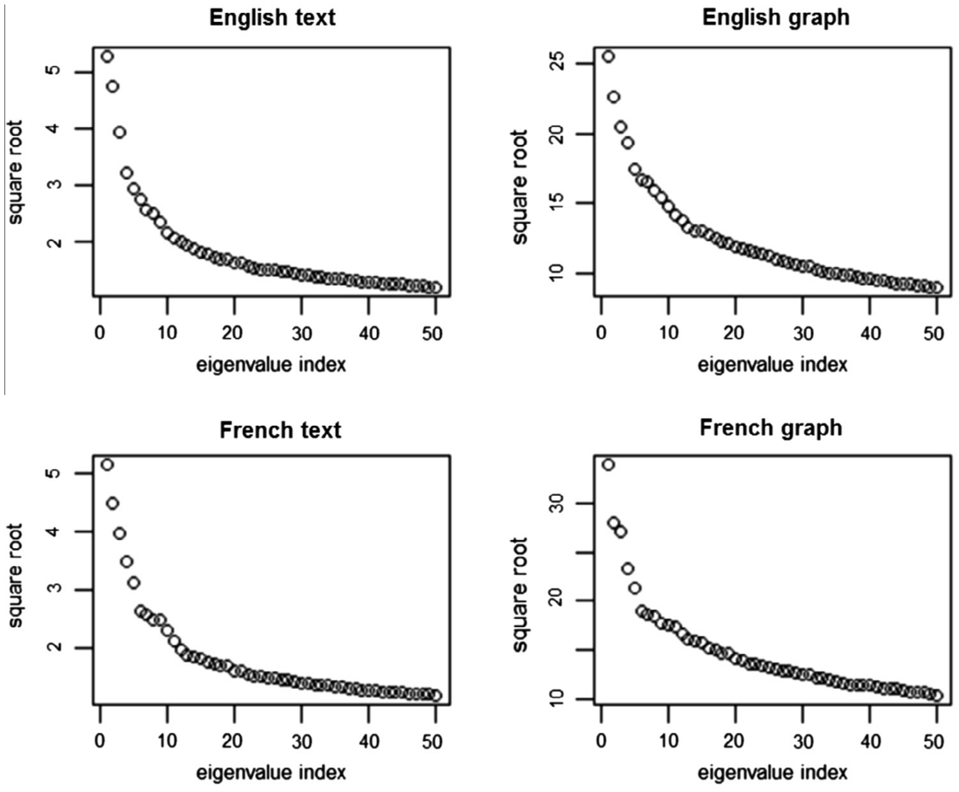


Fig. 4. Squre root of eigenvalues for embedding's variance matrix (all data used).

sults, as described in Section 4.5. These results are illustration of performance to be expected when incorporating a proper model selection methodology.

For canonical correlation analysis, since it requires to multidimensional scale the dissimilarity matrices to d' at the beginning, as described in Section 3.2, when we choose different number n' of domain relation learning training documents, d' depends on n'. We choose the value of d' as large as possible while avoiding numerical underflow. The values of d' with different n' are shown in Table 1. The second column indicates what percentage of the total manifold matching training data $\mathbf{x}_{ik}$ is used.

4.4. Classification setting

The classifiers used in the experiment are κ-nearest neighbor (κ-NN) (Shakhnarovish et al., 2005) and support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000). For κ-NN, the class label of the test data is assigned by the majority class label of the κ closest training data points. The distance used is the usual Euclidean distance. For our experiments we use five-nearest neighbor classifier. SVM sets the separation hyperplane via maximizing the margin. By using the kernel method, SVM can provide non-linear discriminates. We use a polynomial kernel with degree 2.

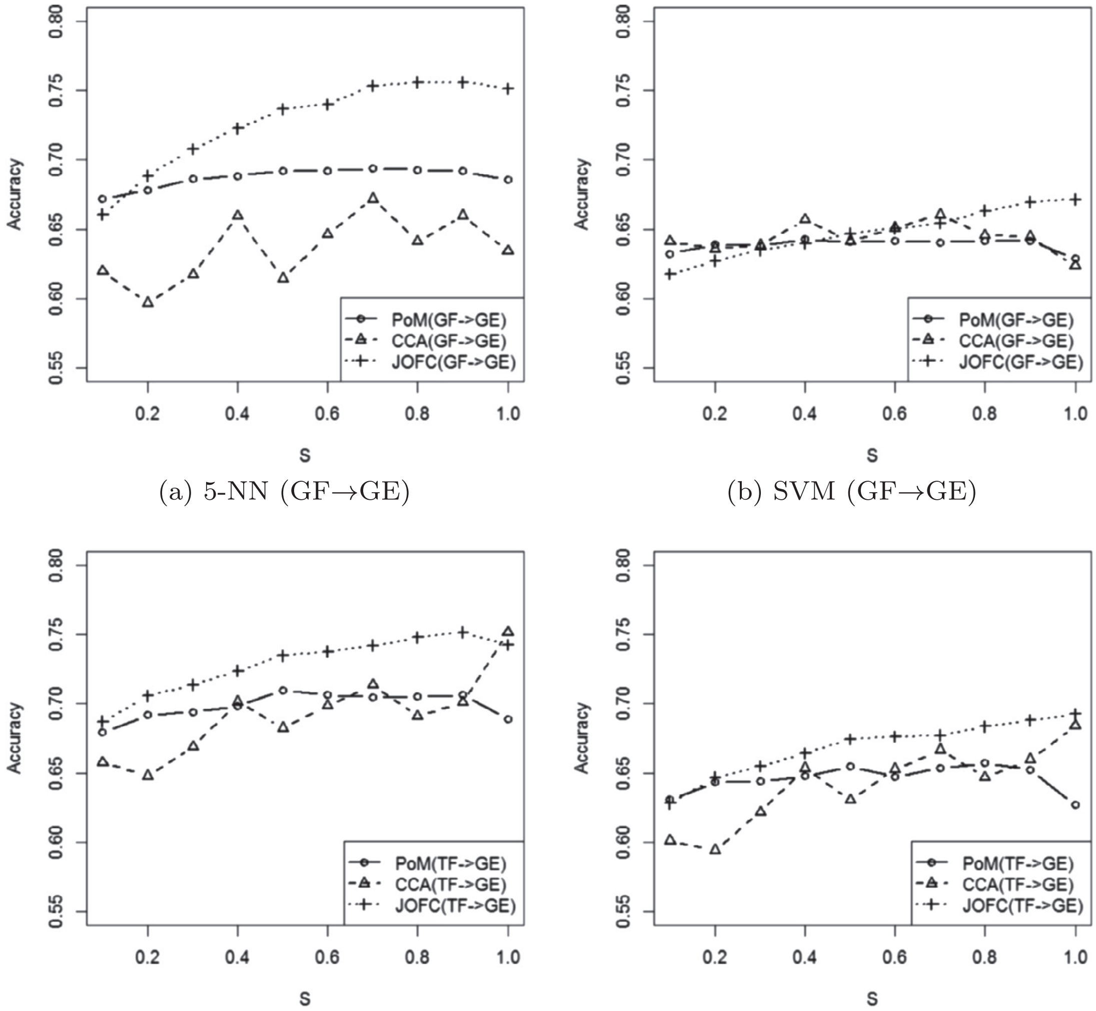
There are 563 new data points $\mathbf{y}_{ik}$ in classes 1 and 3. Class 1 has 372 data points, and the remaining 191 have class label 3. For each n' in Table 1, we randomly sample n' out of the total 819 domain relation learning training documents to learn the common space $\mathbb{R}^d$ into which we project the new data points. The classification is run in a leave-one-out way. We use 200 Monte Carlo replicates to calculate the average performance.

The method described in Section 3.2 generates the embeddings $\tilde{\mathbf{y}}_{ik} \in \mathbb{R}^{15}, i = 1, \ldots, 563, k = 1, 2$. Because there are two kinds of dissimilarity matrices considered, we have $\Delta_k^t \mapsto \tilde{\mathbf{y}}_{ik}^t$ and $\Delta_k^g \mapsto \tilde{\mathbf{y}}_{ik}^g$. The training and testing data can be chosen from not only different spaces (i.e. English space and French space), but also from different dissimilarity measures (i.e. text content dissimilarity and graph topology dissimilarity). Classification results are shown in Fig. 5.

4.5. Classification results

In Fig. 5, the x-axis label S indicates what proportion of the total n data points are used for domain relation learning training, that is, $S = \frac{n'}{n}$; the y-axis is classification accuracy. GE means the embeddings from English graph topology dissimilarity matrix Δ_1^g are used. Similarly, GF and TF represent the embeddings from French graph topology dissimilarity matrix Δ_2^g and French text content dissimilarity matrix Δ_2^t respectively. GF→GE means Δ_2^g is used for classifier training and Δ_1^g is for testing. TF→GE means the classifier is trained on Δ_2^t and tested on Δ_1^g.

In Fig. 5a, Δ_2^g is used for training and Δ_1^g is for testing, thus $\mathbf{x}_{ik}^g, i = 1, \ldots, n', k = 1, 2$ are employed to learn the manifold matching methods. The solid circle curve is for PoM, while the dashed tri-



(a) 5-NN (GF→GE)

(b) SVM (GF→GE)

(c) 5-NN (TF→GE)

(d) SVM (TF→GE)

Fig. 5. Classification accuracy (French graph and text embeding to classify English graph embeding).

Table 2
Classification accuracy.

	GF → GE		TF → GE	
	$S = 10\%$ $d' = 40$	$S = 100\%$ $d' = 200$	$S = 10\%$ $d' = 40$	$S = 100\%$ $d' = 200$
5-NN				
PoM	$67.20\% \pm 0.11\%$	$68.56\% \pm 0.08\%$	$67.92\% \pm 0.14\%$	$68.92\% \pm 0.13\%$
CCA	$61.95\% \pm 0.12\%$	$63.48\% \pm 0.10\%$	$65.75\% \pm 0.12\%$	$75.13\% \pm 0.10\%$
JOFC	$66.08\% \pm 0.15\%$	$75.13\% \pm 0.11\%$	$68.72\% \pm 0.14\%$	$74.25\% \pm 0.10\%$
SVM				
PoM	$63.23\% \pm 0.10\%$	$62.88\% \pm 0.06\%$	$63.10\% \pm 0.15\%$	$62.70\% \pm 0.14\%$
CCA	$64.10\% \pm 0.07\%$	$62.34\% \pm 0.06\%$	$60.09\% \pm 0.12\%$	$68.38\% \pm 0.08\%$
JOFC	$61.76\% \pm 0.12\%$	$67.14\% \pm 0.10\%$	$62.81\% \pm 0.14\%$	$69.27\% \pm 0.12\%$

angle and dotted plus curves represent CCA and JOFC respectively. For each test data point $\tilde{\mathbf{y}}^g_{i1}, i \in \{1, \ldots, m\}$, the 5-NN classifier is trained on $\tilde{\mathbf{y}}^g_{i'2}, i' = 1, \ldots, i-1, i+1, \ldots, m$, and the classification accuracy is calculated as m'/m, where m' is the number of correctly classified testing data points. For each n', 200 Monte Carlo replicates are run to randomly sample n' out of the total n domain relation learning training data points $\mathbf{x}^g_{ik}, i = 1, \ldots, n$. The average accuracy is plotted; the standard errors are available via bootstrap resampling.

Fig. 5c is similar to Fig. 5a except the training data is from Δ^t_2 instead of Δ^g_2. Since Δ^t_2 and Δ^g_1 are within different ranges, prescaling is needed, which is done by $\Delta^t_2 = \Delta^t_2 \frac{\|\Delta^g_1\|_F}{\|\Delta^t_2\|_F}$.

Similarly, Fig. 5b and d show the classification results of PoM, CCA and JOFC using SVM with degree 2 polynomial kernel. Fig. 5b uses Δ^g_2 to classify Δ^g_1, while Fig. 5d uses Δ^t_2 to classify Δ^g_1.

Based on the results shown in Fig. 5a–d, we can see as a general guideline JOFC outperforms both PoM and CCA with regard to the cross-language text document classification. The superior performance of JOFC comes from its ability to jointly preserve fidelity and commensurability in the mapping. We do not claim that this dominance holds uniformly. Indeed, there are exception points in the plots of Fig. 5; for example, when there are few domain relation learning training data (Fig. 5a and b), or for the case when all domain relation learning training data are used (Fig. 5c). But as a general guideline JOFC is a better choice compared to PoM and CCA in terms of classification performance and efficiency.

With increasing amount of domain relation learning training data, the classification performance of JOFC improves, while for PoM and CCA, their classification performance does not necessarily increase with more domain relation learning training data, as shown in Fig. 5a.

In the case of using Δ^g_2 to classify Δ^g_1, for both PoM and JOFC, 5-NN has a higher classification accuracy than SVM with degree 2 polynomial kernel. But for CCA, 5-NN gets lower classification accuracy for certain cases. In the case of using Δ^t_2 to classify Δ^g_1, for all PoM, CCA and JOFC, 5-NN yields better classification performance than SVM.

Table 2 shows the classification accuracy of various methods for $S = 10\%$ and $S = 100\%$. The standard error is obtained via bootstrapping for 1000 samples.

5. Conclusion

In this paper we investigate the performance of three manifold matching methods (PoM, CCA and JOFC) on a cross-language text classification task. We show their performance with manifold matching training data from different domains and different dissimilarity measures, and we also investigate their efficiency by choosing different amounts of domain relation learning training data. In our framework each document is assigned a single topic. The case of multi-topic document assignment from Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) is an interesting extension for ongoing investigation. The experimental results indicate that JOFC, which jointly optimizes fidelity and commensurability, outperforms both PoM and CCA. These results provide significant impetus for further investigation of jointly optimizing fidelity and commensurability for general cross-language inference.

In (Sun et al., 2013), a regularized version of CCA was investigated for the same classification task; the results presented there demonstrate that JOFC is superior to regularized CCA for *GF* → *GE*, but the regularized CCA is superior for *TF* → *GE*. The specific comparative analysis of JOFC vs regularized CCA remains a topic of investigation.

References

Anderson, M.J., Robinson, J., 2003. Generalized discriminant analysis based on distances. Australian & New Zealand Journal of Statistics 45, 301–318.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (6), 1373–1396.

Borg, I., Groenen, P., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer, Verlag.

Cox, T., Cox, M., 2001. Multidimensional Scaling. Chapman and Hall.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines. Cambridge University.

Diaz, F., Metzler, D., 2007. Pseudo-aligned multilingual corpora. In: Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI).

Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K., 1997. Automatic cross-language retrieval using latent semantic indexing. In: AAAI Symposium on Cross Language Text and Speech Retrieval.

Fortuna, B., Shawe-Taylor, J., 2005. The use of machine translation tools for cross-lingual text mining. In: Proceedings of the ICML Workshop on Learning with Multiple Views.

Ham, J., Lee, D., Saul, L., 2005. Semisupervised alignment of manifolds. In: Proc. of the Tenth Int'l Workshop on Artificial Intelligence and Statistics.

Ham, J., Ahn, I., Lee, D., 2006. Learning a manifold-constrained map between image sets: applications to matching and pose estimation. In: Proc, IEE Conf. Computer Vision and Pattern Recognition.

Hardoon, D.R., Szedmak, S.R., Shawe-taylor, J.R., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Computation 16 (12), 2639.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer, Berlin.

Karakos, D., Eisner, J., Khudanpur, S., Priebe, C.E., 2007. Cross-instance tuning of unsupervised document clustering algorithms. In: Proceedings of the Main Conference Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

Kettenring, J.R., 1971. Canonical analysis of several sets of variables. Biometrika 58, 433–451.

Lafon, S., Keller, Y., Coifman, R., 2006. Data fusion and multicue data matching by diffusion maps. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (11), 1784–1797.

Li, Y., Taylor, J.S., 2007. Advanced learning algorithms for cross-language patent retrieval and classification. Information Processing and Management 43 (5), 1183–1199.

Ling, X., Xue, G., Dai, W., Jiang, Y., Yang, Q., Yu, Y., 2008. Can chinese webpages be classified with english data source? In: Proceedings of WWW-08, Beijing, pp. 969–978.

Lu, J., Tan, Y., Wang, G., 2011. Discriminative multi-manifold analysis for face recognition from a single training sample per person. In: IEEE International Conference on Computer Vision (ICCV), pp. 1943–1950.

Ma, Z., Marchette, D., Priebe, C.E., 2012. Fusion and inference from multiple data sources in a commensurate space. Statistical Analysis and Data Mining 5 (3), 187–193.

Mika, S., Schölkopf, B., Smola, A., Müller, K.R., Scholz, M., Rätsch, G., 1999. Kernel PCA and de-noising in feature spaces. Advances in Neural Information Processing Systems 11 (1), 536–542.

Olsson, J.S., Oard, D.W., Hajič, J., 2005. Cross-language text classification. In: Proceedings of SIGIR-05, Salvador, pp. 645–646.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22 (10), 1345–1359.

Pei, Y., Huang, F., Shi, F., Zha, H., 2012. Unsupervised image matching based on manifold alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (8), 1658–1664.

Priebe, C.E., Marchette, D.J., Ma, Z., Adali, S., in press. Manifold matching: joint optimization of fidelity and commensurability. Brazilian Journal of Probability and Statistics.

Rigutini, L., Maggini M., Liu, B., 2005. An em based training algorithm for cross-language text categorization. In: Proceddings of WI05, Compiégne, pp. 529–535.

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5), 513–523.

Shakhnarovish, G., Darrell, T., Indyk, P., 2005. Nearest-Neighbor Methods in Learning and Vision. MIT Press.

Sun, M., Priebe, C.E., Tang, M., 2013. Generalized canonical correlation analysis for disparate data fusion. Pattern Recognition Letters 34 (2), 194–200.

Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for non-linear dimensionality reduction. Science 290, 2319–2323.

Torgerson, W., 1952. Multidimensional scaling: I. Theory and method. Psychometrika.

Trosset, M.W., Priebe, C.E., 2008. The out-of-sample problem for classical multidimensional scaling. Computational Statistics & Data Analysis 52 (10), 4635–4642.

Verbeek, J., Vlassis, N., 2006. Gaussian fields for semi-supervised regression and correspondence learning. Pattern Recognition 39, 1864–1875.

Wang, C., Mahadevan, S., 2008. Manifold alignment using procrustes analysis. In: Proc. Int'l Conf. on Machine Learning, pp. 1120–1127.

Wang, C., Mahadevan, S., 2009. Manifold alignment without correspondence. In: Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 1273–1278.

Wang, R., Chen, X., 2009. Manifold discriminant analysis. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 429–436.

Xiong, L., Wang, F., Zhang, C., 2007. Semi-definite manifold alignment. In: European Conference on Machine Learning, pp. 773–781.