Fusion and Inference from Multiple Data Sources in a Commensurate Space

Zhiliang Ma[1], **David J. Marchette**[2] **and Carey E. Priebe**[1*]

[1]*Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD, USA*

[2]*Naval Surface Warfare Center, Dahlgren, VA, USA*

Received 14 September 2011; revised 2 January 2012; accepted 29 January 2012
DOI:10.1002/sam.11142
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Given objects measured under multiple conditions—for example, indoor lighting versus outdoor lighting for face recognition, multiple language translation for document matching, etc.—the challenging task is to perform data fusion and utilize all the available information for inferential purposes. We consider two exploitation tasks: (i) how to determine whether a set of feature vectors represent a single object measured under different conditions; and (ii) how to create a classifier based on training data from one condition in order to classify objects measured under other conditions. The key to both problems is to transform data from multiple conditions into one commensurate space, where the (transformed) feature vectors are comparable and would be treated as if they were collected under the same condition. Toward this end, we studied Procrustes analysis and developed a new approach, which uses the interpoint dissimilarities for each condition. We impute the dissimilarities between measurements of different conditions to create one omnibus dissimilarity matrix, which is then embedded into Euclidean space. We illustrate our methodology on English and French documents collected from Wikipedia, demonstrating superior performance compared to that obtained via standard Procrustes transformation. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2012

Keywords: fusion; inference; dissimilarity; multidimensional scaling; Procrustes transformation; embedding

1. INTRODUCTION

Information fusion techniques aim to merge information from multiple data sources in order to achieve more accurate inferences than using each single source alone. Information fusion is a relevant and important research field with many applications, such as face recognition, multimodal biometrics, image analysis, and multimedia information retrieval [1–4]. A follow-up work [5] studies the fusion problems in a more theoretical way. The authors put the P- and W-approaches under the same framework and demonstrate that they are two special cases of a general approach.

In general, the most often used information fusion approaches can be summarized into two categories: feature level fusion and decision level fusion. In feature level fusion, feature vectors extracted from different data sources are combined into the Cartesian product space, directly [6] or via some data transformation procedures [3]. Decision level fusion involves combining results obtained separately from all data sources. An ensemble of classifiers is one such example, as is track fusion [7]. The advantage of these two types of information fusion stems from the fact that multiple sets of feature vectors extracted from the same collection of objects usually reflect different characteristics of patterns. By fusing multiple disparate data sources, one generates a more complete representation of the space where the objects live, and hence has more power for inferential tasks such as hypothesis testing, classification, etc.

In this work, we consider information fusion from a different perspective. Suppose that objects are measured under multiple conditions—for example, indoor lighting versus outdoor lighting for face recognition, multiple language translation for document matching, etc. The challenging questions are: (i) how to determine whether a set of feature vectors represents a single object measured under different conditions? For example, whether pictures taken under different lighting conditions are of the same individual; and (ii) how to create a classifier based on training data measured under one condition in order to classify objects measured under other conditions? We refer to the two problems as the implicit translation problem and the classification problem, respectively. A direct approach would involve finding the underlying mappings among all the spaces of measurements and transform these measurements into one commensurate

Correspondence to: cep@jhu.edu

space through the derived mappings. In this commensurate space, all transformed feature vectors are treated equally as if they were from the same data source. The solutions to both questions will then be straightforward. In real applications, finding these mappings is usually difficult. In fact, it is possible to fuse multiple spaces into one commensurate space without the mappings among these spaces. (Generalized) Procrustes analysis is one potential solution. Consider a set of objects, each of which is measured under K ($K \geq 2$) conditions, yielding K feature vectors. Assuming all the feature vectors have been column centered, Procrustes solution rotates (possibly with dilation) the feature vectors to best match each other, and thereby defines a commensurate space. For example, let $\mathbf{X}_0$ and $\mathbf{X}_1$ be two column-centered matrices of the same size, the Procrustes problem is to find a scalar s^* and an orthogonal matrix $\mathbf{Q}^*$ such that $\mathbf{X}_0 \approx s^*\mathbf{X}_1\mathbf{Q}^*$. That is,

$$(s^*, \mathbf{Q}^*) = \underset{\mathbf{Q}\mathbf{Q}^t=\mathbf{I}}{\arg\min} \|\mathbf{X}_0 - s\mathbf{X}_1\mathbf{Q}\|_{\mathrm{F}},$$

where $\|\cdot\|_{\mathrm{F}}$ denote the Frobenius norm.

We use a collection of Wikipedia documents to illustrate the two problems (implicit translation and classification) and the solutions. Because the data are high-dimensional, dissimilarity analysis is used to find a low-dimensional representation. The two step approach, which we refer to as the P-approach, first embeds dissimilarity matrices derived from different data sources and then utilizes a Procrustes transformation on the embeddings to make them commensurate. We propose an approach that simultaneously embeds all dissimilarity matrices and finds the commensurate space. In this approach, dissimilarity matrices from different data sources are put onto the diagonal of an omnibus matrix, whose off-diagonal entries are imputed. Embedding this omnibus matrix results in feature vectors in one commensurate space. We refer to this approach as the W-approach. Both approaches are studied in this work, and the results on Wikipedia example show that the W-approach leads to larger powers in testing and higher accuracy in classification, compared to the P-approach.

In Section 2, we describe the Wikipedia data set, the derivation of dissimilarity matrices, and the implicit transformation and classification problems. Section 3 details the traditional Procrustes solution and the proposed W-approach. We assessed the approaches using simulations, and the results are given in Section 4. Section 5 provides conclusions.

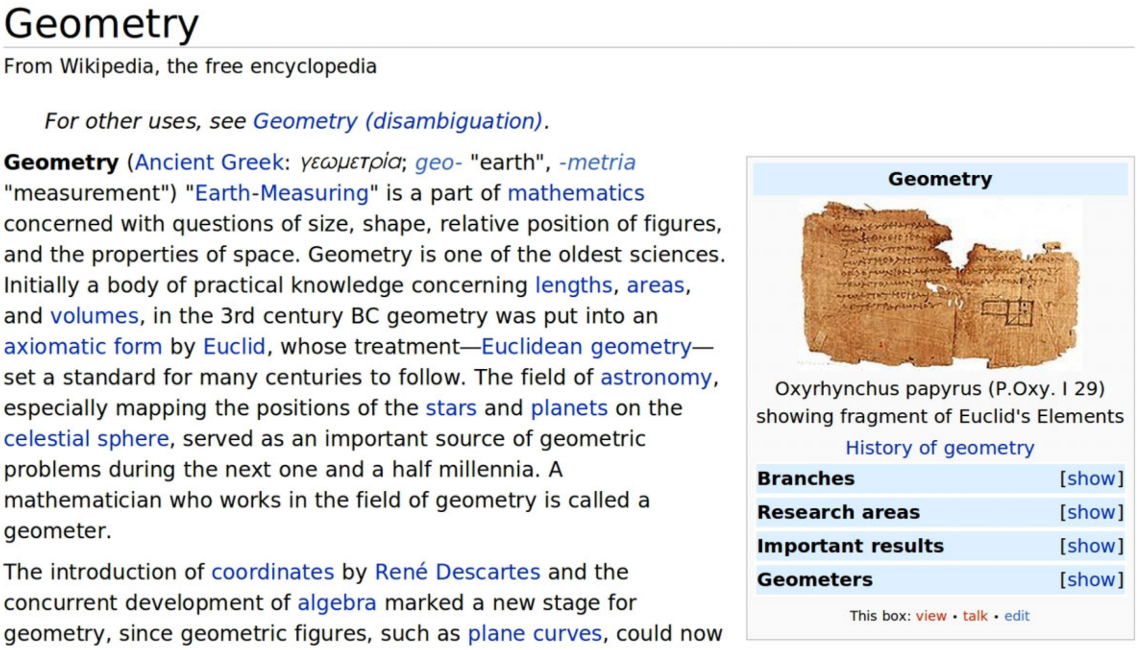
2. DATA

Wikipedia [8] is an open-source Encyclopedia that is written by a large community of users (everyone who wants to, basically). There are versions in over 200 languages, with various amounts of content. The full data for the Wikipediae are freely available for download. A Wikipedia document has one or more of: title, unique ID number, text—the content of the document, images, internal links—links to other Wikipedia documents, external links—links to other content elsewhere on the web, and language links—links to 'the same' document in other languages. Figure 1 shows an English Wikipedia document titled 'Geometry'. The multilingual Wikipediae provide a good testbed for developing methods for analysis of text, implicit translation, and fusion of text and graph information.

2.1. Dissimilarities from Graph Structure and Textual Content

We represent a collection of associated Wikipedia documents as a graph, where nodes correspond to documents and edges denote links among documents. Although the links are directed, for simplicity we treat the resulted graph as an undirected one. We consider two Wikipediae, English and French. A subset is extracted such that there is an one-to-one correspondence between English documents and French documents. We define the 1-neighborhood of a document as the document itself and the documents that have links to or from it. Accordingly, the 2-neighborhood of a document includes its 1-neighborhood, as well as the documents that have links to or from its 1-neighborhood documents. For simplicity, we further reduce the English subset by considering only the 2-neighborhood of the document 'Algebraic Geometry', yielding set $\boldsymbol{E} = \{\boldsymbol{x}_{1,0}, \ldots, \boldsymbol{x}_{n,0}\}$. The set $\boldsymbol{E}$ contains $n = 1382$ English Wikipedia documents. The associated documents in French Wikipedia constitute the set $\boldsymbol{F} = \{\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{n,1}\}$. Thus, the English graph with nodes in $\boldsymbol{E}$ is connected by construction, but the French graph with nodes in $\boldsymbol{F}$ need not be connected (and in fact it is not). We consider two types of data, both of which are given in the form of dissimilarity matrices denoted generically as $\mathbf{D}_0$ and $\mathbf{D}_1$: (i) dissimilarity matrices $\mathbf{G}_0$ and $\mathbf{G}_1$, developed from the graph structures of $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively; (ii) dissimilarity matrices $\mathbf{T}_0$ and $\mathbf{T}_1$, obtained from the textual contents of $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively.

To get dissimilarity matrices from graph structure, the adjacency matrices $\mathbf{A}_0$ and $\mathbf{A}_1$ are first created from $\boldsymbol{E}$ and $\boldsymbol{F}$. An adjacency matrix is a square binary matrix, with 1 in position (i, j) only when the ith document contains a link to or from the jth document. Dissimilarity matrices $\mathbf{G}_0$ and $\mathbf{G}_1$ are then derived from $\mathbf{A}_0$ and $\mathbf{A}_1$, with (i, j) entry as the number of steps it takes to reach node j from node i. By the nature of the graphs, the elements of $\mathbf{G}_0$ take values in $\{0, 1, 2, 3, 4\}$. For example, consider four English Wikipedia documents, $\boldsymbol{x}_{i,0}$, $\boldsymbol{x}_{j,0}$, $\boldsymbol{x}_{r,0}$ and $\boldsymbol{x}_{s,0}$.



Geometry

From Wikipedia, the free encyclopedia

For other uses, see Geometry (disambiguation).

Geometry (Ancient Greek: *γεωμετρία*; *geo-* "earth", *-metria* "measurement") "Earth-Measuring" is a part of mathematics concerned with questions of size, shape, relative position of figures, and the properties of space. Geometry is one of the oldest sciences. Initially a body of practical knowledge concerning lengths, areas, and volumes, in the 3rd century BC geometry was put into an axiomatic form by Euclid, whose treatment—Euclidean geometry—set a standard for many centuries to follow. The field of astronomy, especially mapping the positions of the stars and planets on the celestial sphere, served as an important source of geometric problems during the next one and a half millennia. A mathematician who works in the field of geometry is called a geometer.

The introduction of coordinates by René Descartes and the concurrent development of algebra marked a new stage for geometry, since geometric figures, such as plane curves, could now

Geometry

Oxyrhynchus papyrus (P.Oxy. I 29) showing fragment of Euclid's Elements

History of geometry

Branches [show]

Research areas [show]

Important results [show]

Geometers [show]

This box: view • talk • edit

Fig. 1 ‘Geometry’, an example of English Wikipedia documents. In general, a Wikipedia document has one or more of: title, unique ID number, text, images, internal links, external links, and language links. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Suppose they have the following association with no other links among them:

$$\boldsymbol{x}_{i,0} — \boldsymbol{x}_{j,0} — \text{‘Algebraic Geometry’} — \boldsymbol{x}_{r,0} — \boldsymbol{x}_{s,0}.$$

Then the dissimilarities from $\boldsymbol{x}_{i,0}$ to $\boldsymbol{x}_{j,0}$, $\boldsymbol{x}_{r,0}$ and $\boldsymbol{x}_{s,0}$ are 1, 3 and 4, respectively. Similarly, the elements of $\mathbf{G}_1$ take values in $\{0, 1, \ldots, 1384\}$, with 1384 (by definition) meaning no path between the two corresponding documents—recall that the French graph is not connected. Because it is too computationally intensive to develop $\mathbf{G}_1$, in practice we assign the value 6 to $\mathbf{G}_1(i, j)$ if the shortest path between document i and document j contains more than 4 links. Therefore, the elements of $\mathbf{G}_1$ take values in $\{0, 1, 2, 3, 4, 6\}$.

For dissimilarity matrices of textual content, we use Lin and Pantel's approach [9,10] on Wikipedia documents $\boldsymbol{E}$ and $\boldsymbol{F}$ to obtain two mutual information feature matrices, $\mathbf{MI}_0$ and $\mathbf{MI}_1$. The matrix $\mathbf{MI}_k$ is of size $n \times f_k$, where f_k is the number of features extracted from $\boldsymbol{E}$ (or $\boldsymbol{F}$). Each of the features is associated with a word (after stemming and removal of stopper words). Rare-word discounting [9] is then applied to reduce the impact of infrequent words, yielding $\mathbf{MI}'_0$ and $\mathbf{MI}'_1$. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two rows of $\mathbf{MI}'_k$—that is, they are the feature vectors of two documents. The dissimilarity function ρ is defined as $\rho(\boldsymbol{a}, \boldsymbol{b}) = 1 - (\boldsymbol{a} \cdot \boldsymbol{b})/(\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2)$. Employing ρ separately on the two feature matrices, $\mathbf{MI}'_0$ and $\mathbf{MI}'_1$, results in two dissimilarity matrices $\mathbf{T}_0$ and $\mathbf{T}_1$.

Therefore the Wikipedia data set contains four dissimilarity matrices $\mathbf{G}_0, \mathbf{G}_1$, $\mathbf{T}_0$ and $\mathbf{T}_1$. When a new English document $\boldsymbol{y}_0$ and a new French $\boldsymbol{y}_1$ are provided, we have access to the dissimilarities (for both graph structure and textual content) between $\boldsymbol{y}_0$ and $\boldsymbol{x}_{i,0}$, and those between $\boldsymbol{y}_1$ and $\boldsymbol{x}_{i,1}$, $i = 1, \ldots, n$.

2.2. Implicit Translation and Classification

An implicit translation of a document, unlike a word-level or a real translation in any normal sense, is an association with another document in a different language that is on the same topic. In our framework, we treat each topic as an object with measurements (documents) taken under different conditions (languages). We say that documents of different languages are matched if they are on the same topic. For example, the English Wikipedia document ‘Standard deviation’ and the French Wikipedia document with the same title are two matched documents. We represent two matched documents, for example, an English document $\boldsymbol{x}_{i,0}$ and a French document $\boldsymbol{x}_{i,1}$, as $\boldsymbol{x}_{i,0} \sim \boldsymbol{x}_{i,1}$. The goal of implicit translation is to determine whether a match is present between two new documents $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$. That is, we consider the hypothesis testing:

$$H_0 : \boldsymbol{y}_0 \sim \boldsymbol{y}_1 \quad \text{versus} \quad H_A : \boldsymbol{y}_0 \nsim \boldsymbol{y}_1.$$

Notice that we assume the two new documents represent a matched pair under H_0. This allows us to control the probability of missing a true match. This is practical in many applications where computer algorithms are used to eliminate easily rejected pairs and the remaining possibly matched pairs will be manually examined.

In the classification problem, we have a collection of one-to-one matched English and French Wikipedia documents, which have been classified into one of the classes in $J = \{0, 1, \ldots\}$. Some classified English Wikipedia documents with class labels in $\tilde{J} = \{|J| + 1, |J| + 2, \ldots\}$ are also available ($|J|$ denotes the size of J). We are interested in classifying new French Wikipedia documents whose classes are in $\tilde{J}$ (but which classes are unknown). Formally, consider two manifolds, Ξ_0 and Ξ_1. Let

$$\begin{aligned}
&(X, C, Z) \sim F_{X,C,Z},\\
&C : \Omega \to J \cup \tilde{J},\\
&Z : \Omega \to \{0, 1\},\\
&X|Z = z : \Omega \to \Xi_z,
\end{aligned}$$

where J and $\tilde{J}$ are two disjoint sets of class labels. Suppose the following training data are available

$$\begin{aligned}
\mathcal{T}_0 &= \{(x_i, c_i \in J, z_i = 0),\ i = 1, \ldots, n_0\},\\
\mathcal{T}_1 &= \{(x_i, c_i \in J, z_i = 1),\ i = 1, \ldots, n_1\},\\
\tilde{\mathcal{T}}_0 &= \{(x_i, c_i \in \tilde{J}, z_i = 0),\ i = 1, \ldots, m_0\},
\end{aligned}$$

where n_0, n_1, and m_0 are the number of observations in the sets $\mathcal{T}_0$, $\mathcal{T}_1$, and $\tilde{\mathcal{T}}_0$, respectively. We are interested in creating a classifier g based on the training data and use it to classify future observations in Ξ_1 into one of the classes in $\tilde{J}$. In the Wikipedia example, $\mathcal{T}_0$ and $\tilde{\mathcal{T}}_0$ denote the training English documents with labels in $J \cup \tilde{J}$, and $\mathcal{T}_1$ denotes the training French documents with labels in J. Figure 2 depicts the classification problem.

We consider Ξ_0 and Ξ_1 to be the English and French Wikipedia document space, respectively. The 1382 Wikipedia documents are labeled into five groups. The two disjoint sets of class labels are $J = \{0, 1, 2\}$ and $\tilde{J} = \{3, 4\}$. We are interested in finding a way to create a classifier based on English documents and use it to classify French documents.

3. METHODS

For the implicit translation problem, suppose that there is a way to embed $\boldsymbol{E} \in \Xi_0$ and $\boldsymbol{F} \in \Xi_1$ into a commensurate space Ξ_c, where the embeddings of English and French documents would be treated as if they were collected under the same condition. We can embed the two new documents $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$, referred to as the out-of-sample documents, into the space Ξ_c. Whether a match is present is then determined by examining the Euclidean distance between the embeddings of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$, with a large distance being evidence against H_0. There are two ways to determine critical values. The naïve way is to treat the distances between the embeddings of matched pairs in $\mathbf{E}$ and $\mathbf{F}$ as the ground truth, and use the $100(1 - \alpha)$th percentile as the critical value for a level α test. However, this method does not always lead to large powers, because the distribution of the distances between out-of-sample embeddings is usually slightly different from that of the original embeddings, even under the matched assumption H_0. Another way of obtaining critical values is by means of Monte Carlo simulation: (i) randomly choose a pair of matched documents $\boldsymbol{x}_{i,0}$ and $\boldsymbol{x}_{i,1}$ from $\boldsymbol{E}$ and $\boldsymbol{F}$, and treat them as out-of-sample documents; (ii) embed the selected documents into the space Ξ_c, and compute their distance; and (iii) repeat (i–ii) to obtain an empirical distribution of such distances. The critical value for a level α test is then calculated as the $100(1 - \alpha)$th percentile of this empirical distribution. We use the latter method in this work and get larger powers than using the naïve approach.

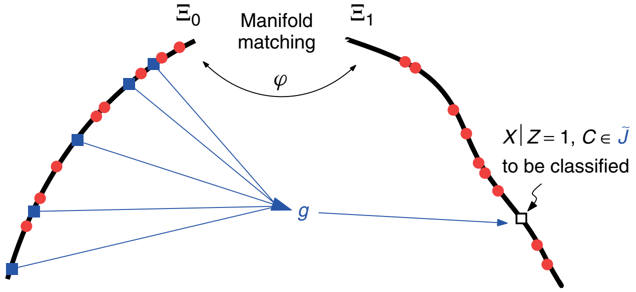


Fig. 2 Classification problem. In space Ξ_0 training data from classes J (red disk) and $\tilde{J}$ (blue square) are available, while in space Ξ_1 only training data from classes J are available. We are interested in training a rule g to classify objects of classes $\tilde{J}$ in space Ξ_1. It is impossible to directly create such a classifier in Ξ_1 due to lack of training data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For the classification problem, suppose a commensurate space Ξ_c could be obtained through $\mathcal{T}_0$ and $\mathcal{T}_1$—English and French documents of classes in J. We can embed the training English documents $\tilde{\mathcal{T}}_0$ and the new French documents into the space Ξ_c. In the commensurate space Ξ_c, building classifier g based on English documents with labels in $\tilde{J}$ and using it to classify new French documents are then straightforward.

Therefore the key to both problems is: how shall we determine the commensurate space Ξ_c and how shall we embed new documents into this space?

3.1. Procrustes Transformation

The Procrustes analysis [11, and references contained therein] is to transform a configuration of points (source) to another (target) as closely as possible in the least-square sense. The permitted transformations are any combination of dilation (uniform scaling), rotation, reflection, and translation. We define the space where the target and the transformed source live as the commensurate space.

For the implicit translation problem, we embed $\mathbf{D}_0$ and $\mathbf{D}_1$ through multidimensional scaling to obtain $n \times d$ configurations $\mathbf{X}_0$ and $\mathbf{X}_1$ in the space $\mathbb{R}^d$ separately. The two new documents $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are then embedded to $\tilde{\boldsymbol{y}}_0$ and $\tilde{\boldsymbol{y}}_1$ in $\mathbb{R}^d$ respectively via out-of-sample embedding [12]. Notice that the coordinates in $\mathbf{X}_0$ and $\mathbf{X}_1$ may be given in different systems. Procrustes analysis is performed to transform one of the embeddings—for example, $\mathbf{X}_1$—to best match the other one—for example, $\mathbf{X}_0$. The resulting transformation function t is then applied to the corresponding out-of-sample embedding $\tilde{\boldsymbol{y}}_1$ so that $t(\tilde{\boldsymbol{y}}_1)$ and $\tilde{\boldsymbol{y}}_0$ are commensurate.

For the classification problem, a similar procedure is performed. Let $\mathbf{D}_0^J$ and $\mathbf{D}_1^J$ denote the dissimilarity matrices among documents in $\mathcal{T}_0$ and $\mathcal{T}_1$. We embed $\mathbf{D}_0^J$ and $\mathbf{D}_1^J$ to obtain $\mathbf{X}_0^J$ and $\mathbf{X}_1^J$ in $\mathbb{R}^d$, respectively. Then the English documents in $\tilde{\mathcal{T}}_0$ and the new French documents, whose class labels belong to $\tilde{J}$ (but which classes are unknown), are embedded via out-of-sample embedding, yielding $X_0^{\tilde{J}}$ and $X_1^{\tilde{J}}$ in $\mathbb{R}^d$. Procrustes transformation function t_J learned from $\mathbf{X}_0^J$ and $\mathbf{X}_1^J$ is then applied to $X_1^{\tilde{J}}$ so that $t_J(X_1^{\tilde{J}})$ and $X_0^{\tilde{J}}$ are commensurate.

We refer this approach as the P-approach.

3.2. Our Approach

The P-approach creates a commensurate space in two steps, namely embedding and Procrustes transformation. In this section, we introduce a novel method, which defines a commensurate space in one step. Suppose that we have access to a $2n \times 2n$ dissimilarity matrix, which consists of the pairwise dissimilarities among documents in $\boldsymbol{E} \cup \boldsymbol{F} = \{\boldsymbol{x}_{1,0}, \ldots, \boldsymbol{x}_{n,0},\ \boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{n,1}\}$. Then the embedding of this dissimilarity matrix is a $2n \times d$ data matrix, with the first n rows being the embedding of $\boldsymbol{E}$ and the rest the embedding of $\boldsymbol{F}$. In addition, the embeddings of $\boldsymbol{E}$ and $\boldsymbol{F}$ are in the same space, that is, the commensurate space. The question is how to obtain the $2n \times 2n$ omnibus dissimilarity matrix.

In implicit translation, we impute $\mathbf{W}$, the dissimilarities between $\boldsymbol{E}$ and $\boldsymbol{F}$, by the entrywise average of $\mathbf{D}_0$ and $\mathbf{D}_1$. That is, the dissimilarity between the English document $\boldsymbol{x}_{i,0}$ and the French document $\boldsymbol{x}_{j,1}$ is imputed as the average of the following two dissimilarities: the dissimilarity between the English documents $\boldsymbol{x}_{i,0}$ and $\boldsymbol{x}_{j,0}$, and the dissimilarity between the French documents $\boldsymbol{x}_{i,1}$ and $\boldsymbol{x}_{j,1}$. An omnibus dissimilarity matrix $\mathbf{M}$ is then constructed by putting $\mathbf{D}_0$ and $\mathbf{D}_1$ on the diagonal, and putting $\mathbf{W}$ on the off-diagonal. We embed $\mathbf{M}$ to obtain a configuration of $2n$ points $\mathbf{X}$ in $\mathbb{R}^d$. We take the first n points, $\mathbf{X}_0$, and the remaining n points, $\mathbf{X}_1$, as embeddings of $\mathbf{D}_0$ and $\mathbf{D}_1$, respectively. Notice that $\mathbf{X}_0$ and $\mathbf{X}_1$ are already in the same space Ξ_c, because the dissimilarities between matched English and French document pairs have been taken into account when embedding $\mathbf{M}$—the imputed matrix $\mathbf{W}$ has all zeros on its diagonal. For any two additional documents $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$, let $\boldsymbol{u}_0$ denote the dissimilarity vector between the new English Wikipedia document $\boldsymbol{y}_0$ and the original English Wikipedia documents in $\boldsymbol{E}$, and let $\boldsymbol{v}_1$ denote the dissimilarity vector between the new French Wikipedia document $\boldsymbol{y}_1$ and the original French Wikipedia documents in $\boldsymbol{F}$. Under the null hypothesis that $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are matched, we impute the dissimilarities between $\boldsymbol{y}_0$ and $\boldsymbol{F}$ (denoted by $\boldsymbol{v}_0$), and dissimilarities between $\boldsymbol{y}_1$ and $\boldsymbol{E}$ (denoted by $\boldsymbol{u}_1$) by entrywise average of $\boldsymbol{u}_0$ and $\boldsymbol{v}_1$. That is, $\boldsymbol{v}_0 = \boldsymbol{u}_1 = (\boldsymbol{u}_0 + \boldsymbol{v}_1)/2$. Out-of-sample embedding is used to embed $(\boldsymbol{u}_0^t, \boldsymbol{v}_0^t)^t$ and $(\boldsymbol{u}_1^t, \boldsymbol{v}_1^t)^t$ into Ξ_c. Figure 3 depicts the construction of the omnibus dissimilarity matrix $\mathbf{M}$ and the imputation of dissimilarities related to the out-of-sample observations.

In the classification problem, similarly we create omnibus matrix $\mathbf{M}^J$ from $\mathbf{D}_0^J$, $\mathbf{D}_1^J$ and the imputed matrix $\mathbf{W}^J = (\mathbf{D}_0^J + \mathbf{D}_1^J)/2$. The omnibus matrix $\mathbf{M}^J$ is then embedded into a commensurate space Ξ_c. To embed out-of-sample English documents in $\tilde{\mathcal{T}}_0$, we first impute the dissimilarity between $\boldsymbol{x}_{i,0} \in \tilde{\mathcal{T}}_0$ and $\boldsymbol{x}_{j,1} \in \mathcal{T}_1$ by the average of the dissimilarities between $\boldsymbol{x}_{j,1}$ and $\boldsymbol{x}_{i,0}$'s three nearest neighbors in $\mathcal{T}_0$. (These dissimilarities can be found in $\mathbf{W}^J$.) All the imputed dissimilarities are stored in $\mathbf{D}_{01}^{\tilde{J}J}$. The dissimilarities between documents in $\tilde{\mathcal{T}}_0$ and $\mathcal{T}_0$ are given by $\mathbf{D}_0^{\tilde{J}J}$, and the dissimilarities among $\tilde{\mathcal{T}}_0$ are given by $\mathbf{D}_0^{\tilde{J}}$. Trosset and Priebe's out-of-sample embedding approach [12] is then used to embed $\tilde{\mathcal{T}}_0$ into the space Ξ_c. Similarly, new French documents of classes $\tilde{J}$ are embedded into Ξ_c. Figure 4 depicts the construction of the omnibus dissimilarity matrix $\mathbf{M}^J$ and how to out-of-sample embed documents in $\tilde{\mathcal{T}}_0$.

$$\overset{2n\times 2n}{\mathrm{M}} = \begin{array}{cc} \left[\begin{array}{cc} \overset{n\times n}{\mathrm{D}_0} & \overset{n\times n}{\mathrm{W}} \\ \mathrm{W}^T & \overset{n\times n}{\mathrm{D}_1} \end{array}\right] & \begin{array}{cc} \overset{n\times 1}{u_0} & \overset{n\times 1}{u_1} \\ \overset{n\times 1}{v_0} & \overset{n\times 1}{v_1} \end{array} \\ \begin{array}{cc} u_0^t & v_0^t \\ u_1^t & v_1^t \end{array} & \end{array}$$

Fig. 3 We impute $\mathbf{W}$, the dissimilarities between $\boldsymbol{E}$ and $\boldsymbol{F}$, by $(\mathbf{D}_0 + \mathbf{D}_1)/2$ to construct $\mathbf{M}$, which is then embedded into the space Ξ_c. We impute $\boldsymbol{u}_1$ and $\boldsymbol{v}_0$ by $(\boldsymbol{u}_0 + \boldsymbol{v}_1)/2$. Finally, out-of-sample embedding is used to embed $(\boldsymbol{u}_0^t, \boldsymbol{v}_0^t)^t$ and $(\boldsymbol{u}_1^t, \boldsymbol{v}_1^t)^t$ into Ξ_c.

$$\mathrm{M}^J = \begin{array}{|c|c|c|} \hline \mathrm{D}_0^J & \mathrm{W}^J & \mathrm{D}_0^{J\tilde{J}} \\ \hline \mathrm{W}^J & \mathrm{D}_1^J & \mathrm{D}_{10}^{J\tilde{J}} \\ \hline \mathrm{D}_0^{\tilde{J}J} & \mathrm{D}_{01}^{\tilde{J}J} & \mathrm{D}_0^{\tilde{J}} \\ \hline \end{array}$$

Fig. 4 We impute $\mathbf{W}^J$, the dissimilarities between documents in $\mathcal{T}_0$ and $\mathcal{T}_1$, by $(\mathbf{D}_0^J + \mathbf{D}_1^J)/2$ to construct $\mathbf{M}^J$, which is then embedded into the space Ξ_c. The dissimilarities between documents in $\tilde{\mathcal{T}}_0$ and $\mathcal{T}_0$ are given by $\mathbf{D}_0^{\tilde{J}J}$ ($\mathbf{D}_0^{J\tilde{J}}$ is the transpose of $\mathbf{D}_0^{\tilde{J}J}$). The dissimilarity between $\boldsymbol{x}_{i,0} \in \tilde{\mathcal{T}}_0$ and $\boldsymbol{x}_{j,1} \in \mathcal{T}_1$ are imputed by the average of the $\mathbf{W}^J$ entries that are corresponding to $\boldsymbol{x}_{i,1}$ and $\boldsymbol{x}_{i,0}$'s three nearest neighbors in $\mathcal{T}_0$. All the imputed dissimilarities are stored in $\mathbf{D}_{01}^{\tilde{J}J}$ ($\mathbf{D}_{10}^{J\tilde{J}}$ is the transpose of $\mathbf{D}_{01}^{\tilde{J}J}$).

We refer this approach as the W-approach.

3.3. Fusion

We consider one additional step, to combine the data of textual content and graph structure. Ideally both sources of data contain complementary information so that their fusion leads to larger power in testing and higher accuracy in classification than using either textual content data or graph structure data alone. We achieve the fusion by combining the embeddings obtained in the P- or W-approach via the Cartesian product [6].

4. RESULTS

To compute critical values and estimate powers in hypothesis testing, we randomly select two pairs of matched documents from $\boldsymbol{E}$ and $\boldsymbol{F}$. That is, we leave out four documents, two from each language, and they result in two matched pairs and two nonmatched pairs. (Notice that in a real problem we only need to leave out one matched pair out to get critical values; leaving two matched pairs out makes it also possible to estimate testing powers.) The approaches introduced in Section 3 are then applied to obtain the distances between the two matched pairs (denoted by d_0), and the distances between the two nonmatched pairs (denoted by d_A). We use Classical Multidimensional Scaling (CMDS) [13,14] in the embedding. Embedding dimension $d = 6$ is determined by Zhu and Ghodsi's automatic dimensionality selection [15]. We use ranks of the distances d_A based on 200 Monte Carlo simulations to estimate the powers for different levels of α, where the power β_α is the probability of rejecting the null hypothesis when rejection is in fact the correct decision and α is the probability of missing a true match. That is, for each $\alpha \in [0, 1]$, the critical value c_α is defined as the (100α)th percentile of d_0, and the corresponding power is the percentage of distances in d_A that are larger than the critical value c_α. The power at level α is our performance in determining that a non-match is in fact a non-match. The β against α ROC curves are shown in Figure 5. For example, at $\alpha = 0.05$ (missing 5% of the true matches), we obtain a power of $\hat{\beta}_{W\text{-}fusion} = 0.560$ (correctly eliminating 56% of the false matches) via W-fusion. This is a statistical significant improvement over the results obtained sans fusion ($\hat{\beta}_{P\text{-}G} = 0.135$, $\hat{\beta}_{P\text{-}T} = 0.379$, $\hat{\beta}_{W\text{-}G} = 0.403$, $\hat{\beta}_{W\text{-}T} = 0.468$. See Figure 5).

As mentioned in Section 3, the commensurate space Ξ_c in the classification problem is determined by $\mathbf{D}_0^J$ and $\mathbf{D}_1^J$. Training English documents in $\tilde{\mathcal{T}}_0$ and new French documents are then embedded into Ξ_c. We consider two types of association relations between $\mathcal{T}_0$ and $\mathcal{T}_1$, one-to-one association and group association. When assuming one-to-one association, we use the information of one-to-one correspondence between the training English and French documents with classes in J; while for group association, we use only the class label information between English and French documents, but do not use the one-to-one relationship between them. Introducing group association between $\mathcal{T}_0$ and $\mathcal{T}_1$ makes it possible to define a commensurate space through nonmatched English and

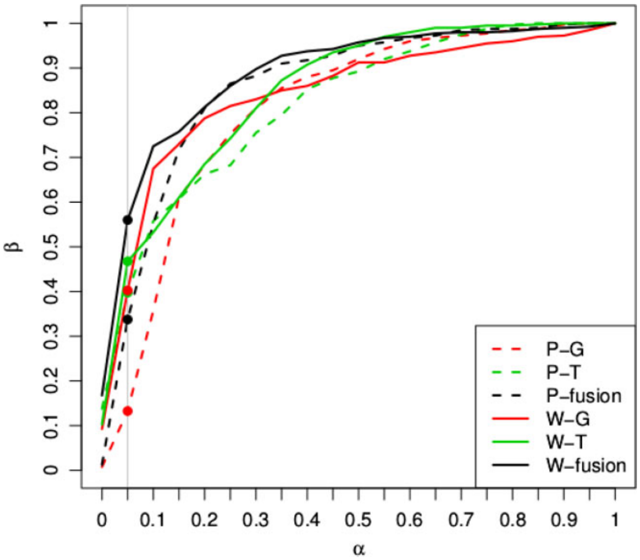


Fig. 5 The ROC curve depicts that W-approach is generally superior to P-approach; T is generally superior to G; Fusion is generally superior to either G or T alone. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 1. Given the association between the training data $\mathcal{T}_0$ and $\mathcal{T}_1$, one-to-one or group-to-group, we transform Ξ_0 and Ξ_1 into one commensurate space by P- or W-approach. A linear discriminant classifier is then created based on $\tilde{\mathcal{T}}_0$ and then tested on $\tilde{\mathcal{T}}_1$. The symbols G and T represent the Graph and Text data, respectively.

Assocation	P-G	P-T	P-fusion	W-G	W-T	W-fusion
One-to-one	0.417	0.496	0.493	0.300	0.285	0.282
Group	0.404	0.470	0.470	0.301	0.069	0.122

French documents. When assuming group association, in the P-approach we learn the transformation matrix through the group means of embeddings. While in the W-approach we impute the dissimilarities among the same group by 0s and those between different groups by the dissimilarities between group means.

In the commensurate space, we train a linear classifier g based on the embedding of $\tilde{\mathcal{T}}_0$. We then apply g to the embeddings of new French documents. Classification errors are given in Table 1. It is clear that the W-approach results in smaller classification errors than the P-approach. However, combining data from the graph structure and the text content does not, in general, improve performance.

5. CONCLUSION

We have discussed two problems regarding fusion from multiple data sources in a commensurate space:

1. how to determine whether a set of feature vectors represent a single object measured under different conditions?

2. how to create a classifier based on training data measured under one condition in order to classify objects measured in other conditions?

The key to both problems is to construct a commensurate space, where the (transformed) feature vectors of different sources are comparable and would be treated as if they were collected under the same condition. Two approaches were studied. In the P-approach, embedding dissimilarity matrices and defining a commensurate space are performed separately. The W-approach achieves the two procedures simultaneously, by constructing an omnibus dissimilarity matrix. Applying both approaches on a Wikipedia data set showed that the W-approach leads to higher hypothesis testing powers in the implicit translation problem and smaller errors in the classification problem, compared to the P-approach.

ACKNOWLEDGMENTS

This work is partially supported by National Security Science and Engineering Faculty Fellowship (NSSEFF), Air Force Office of Scientific Research (AFOSR)—D. J. Marchette was specifically supported by the grant NCRADA-NSWCDD-11-115, Office of Naval Research (ONR), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the American Society for Engineering Education (ASEE) Sabbatical Leave Program.

REFERENCES

[1] C. Liu, and H. Wechsler, A shape- and texture-based enhanced fisher classifier for face recognition, IEEE Trans Image Process 10 (2001), 598–608.

[2] A. Ross, and A. K. Jain, Multimodal biometrics: an overview, In Proceedings of 12th Signal Processing Conference (EUSIPCO), 2004, 1221–1224.

[3] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, A new method of feature fusion and its application in image recognition, Pattern Recognition 38(12) (2005), 2437–2448.

[4] J. Kludas, E. Bruno, and S. M. Maillet, Information Fusion in Multimedia Information Retrieval, Berlin, Heidelberg, Springer-Verlag, 2008.

[5] C. E. Priebe, D. J. Marchette, Z. Ma, S. Adali, Manifold Matching: Joint Optimization of Fidelity and Commensurability, Braz J Prob Stat, accepted for publication, 2012.

[6] Z. Ma, A. Cardinal-Stakenas, Y. Park, M. Trosset, and C. Priebe, Dimensionality Reduction on the Cartesian Product of Embeddings of Multiple Dissimilarity Matrices, J Classif 27 (3) (2010), 307–321.

[7] K. C. Chang, T. Zhi, and R. K. Saha, Performance evaluation of track fusion with information matrix filter, IEEE Trans Aerospace Electron Syst 38 (2002), 455–466.

[8] Wikipedia http://en.wikipedia.org/wiki/Wikipedia. [Last accessed January 2012].

[9] D. Lin, and P. Pantel, Concept discovery from text, In Proceedings of the 19th International Conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics, 2002, 1–7.

[10] P. Pantel, and D. Lin, Discovering word senses from text, In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2002, 613–619.

[11] R. Sibson, Studies in the robustness of multidimensional scaling: Procrustes statistics, J R Stat Soc Ser B 40(2) (1978), 234–238.

[12] M. W. Trosset, and C. E. Priebe, The out-of-sample problem for classical multidimensional scaling, Comput Stat Data Anal 52(10) (2008), 4635–4642.

[13] W. Torgerson, Multidimensional scaling: I. theory and method, Psychometrika 17 (1952), 401–419.

[14] T. F. Cox, and M. A. A. Cox, Multidimensional Scaling, Boca Raton, Chapman & Hall/CRC, 2001.

[15] M. Zhu, and A. Ghodsi, Automatic dimensionality selection from the screen plot via the use of profile likelihood, Comput Stat Data Anal 51 (2006), 918–930.