Pattern Recognition Letters 92 (2017) 41–48

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Manifold matching using shortest-path distance and joint neighborhood selection

Cencheng Shen[a,b,*], Joshua T. Vogelstein[a,c], Carey E. Priebe[a,d,*]

[a] Center for Imaging Science, Johns Hopkins University, USA
[b] Department of Statistics, Temple University, USA
[c] Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, USA
[d] Department of Applied Mathematics and Statistics, Johns Hopkins University, USA

ARTICLE INFO

Article history:
Received 3 September 2016
Available online 3 April 2017

Keywords:
Nonlinear transformation
Seeded graph matching
Geodesic distance
k-nearest-neighbor

ABSTRACT

Matching datasets of multiple modalities has become an important task in data analysis. Existing methods often rely on the embedding and transformation of each single modality without utilizing any correspondence information, which often results in sub-optimal matching performance. In this paper, we propose a nonlinear manifold matching algorithm using shortest-path distance and joint neighborhood selection. Specifically, a joint nearest-neighbor graph is built for all modalities. Then the shortest-path distance within each modality is calculated from the joint neighborhood graph, followed by embedding into and matching in a common low-dimensional Euclidean space. Compared to existing algorithms, our approach exhibits superior performance for matching disparate datasets of multiple modalities.

1. Introduction

The abundance of data in the modern age has made it crucial to effectively deal with large amounts of high-dimensional data. For the purpose of data analysis, it is imperative to apply dimension reduction to embed data into a low-dimensional space for subsequent analysis. Traditional linear embedding techniques have solid theoretical foundations and are widely used, e.g., principal component analysis (PCA) [22,44] and multidimensional scaling (MDS) [5,7,45] for datasets of a single modality, and canonical correlation analysis (CCA) [1,20] for datasets of multiple modalities.

However, real datasets often exhibit nonlinear geometry, discovering which can be advantageous for subsequent inference. Many manifold learning algorithms have been proposed to learn the intrinsic low-dimensional structure of nonlinear datasets, including Isomap [38,43], locally linear embedding (LLE) [30,31], Hessian LLE [9], Laplacian eigenmaps [2,18], local tangent space alignment (LTSA) [50,51], among many others. Most of them start with the assumption that the data are locally linear, explore the local geometry via the nearest-neighbor graph of the sample data, transform the data using the neighborhood graph, and eventually learn the low-dimensional manifold by optimizing some objective function. These nonlinear embedding algorithms usually serve as a preliminary feature extraction step that enables subsequent inference. They have been used successfully in object recognition and image processing.

In this paper, we consider the manifold matching task for datasets of multiple modalities, which is traditionally modeled by multiple dependent random variables. Conventional methods for identifying the relationship among multiple random variables are still very popular in theory and practice, such as canonical correlation [17,20,23] and Procrustes transformation [14,15,36,37]. However, it has become a much more challenging task to match real datasets of multiple modalities from disparate sources due to their complex dependency structures, such as the same document in different languages, an image and its descriptions, or networks of the same actors on different social websites.

There have been many recent endeavors regarding data fusion and manifold matching [25,29,33,35,40,48,49]. Similar to dimension reduction for datasets of a single modality, manifold matching can serve as a feature extraction step to explore datasets of multiple modalities, and has also been shown to help subsequent inference in object recognition [24], information retrieval [39], and transfer learning [28]. Furthermore, the matching task is important on its own and has been applied to explore multiple graphs and networks [26,27,47]. One such application is seeded graph matching, where two large networks are collected but only a percentage of training vertices have known correspondence. Then the remaining vertices need to be properly matched to uncover potential correspondence and benefit later inference.

* Corresponding authors.
E-mail addresses: cshen6@jhu.edu (C. Shen), jovo@jhu.edu (J.T. Vogelstein), cep@jhu.edu (C.E. Priebe).

http://dx.doi.org/10.1016/j.patrec.2017.04.005
0167-8655/

Due to the success of nonlinear embedding algorithms for datasets of a single modality, it is often perceived that these algorithms can be directly combined into the matching framework to improve the matching performance when one or more modalities are nonlinear. A naïve procedure is to pick one nonlinear algorithm, apply it to each modality separately, and match the embedded modalities. But such a simplistic procedure does not always guarantee a good matching performance, since many nonlinear embedding algorithms only preserve the local geometry up to some affine transformation [13]. Furthermore, using nonlinear transformations separately can even deteriorate the matching performance when compared to using simple linear transformations, as shown in our numerical simulations.

To tackle the problem, we propose a manifold matching algorithm using shortest-path distance and joint neighborhood selection. By utilizing a robust distance measure that approximates the geodesic distance, and effectively combining the correspondence information into the embedding step, the proposed algorithm can significantly improve the matching quality from disparate data sources, compared to directly take linear or nonlinear embeddings for matching. All code and data are made publicly available.[1]

2. Manifold matching

In this section, the matching framework and evaluation criteria are first introduced. Next we present the main algorithm, followed by relevant implementation details. Additional discussions are offered on issues that can affect the matching performance.

2.1. The matching framework

Suppose n objects are measured under two different sources. Then $X_l = \{x_{il}\} \in \Xi_l$ for $l = 1, 2$ are the actual datasets that are observed / collected, with $x_{i1} \sim x_{i2}$ for each i ($\sim$ means the two observations are matched in the context). Thus X_1 and X_2 are the two different views / modalities of the same underlying data. This setting is extendable to datasets of more than two modalities, but for ease of presentation we focus mainly on the matching task of two modalities.

Ξ_1 and Ξ_2 are potentially very different from each other, such as a flat manifold and its nonlinear transformation, an image and its description, or texts under different languages. A typical example is the social network, where many users have accounts on Youtube, Facebook, Twitter, etc. People sometimes post different contents and connect with different groups on each network site, such that data analysis of better quality is only possible when multiple accounts of the same person are combined. Some users already linked their accounts from different places, or unique user information are filled (like actual name, occupation), certain accounts can be automatically matched, providing a set of matched training data; but all the other accounts need to be matched by machine (as manual match is too expensive for millions of accounts), presenting a set of testing data from each website.

We assume $x_{il} \in \Xi_l$ is endowed with a distance measure Δ_l such that $\Delta_l(i, j) = dist(x_{il}, x_{jl})$. To match multiple modalities, we find two mappings $\rho_l : \Xi_l \to \mathbb{R}^d, l = 1, 2$ such that the mapped data $\hat{X}_l = \{\rho_l(x_{il})\}$ are matched into a common low-dimensional Euclidean space $\mathbb{R}^d$. A simple example of ρ_l is MDS (e.g., classical MDS first doubly centers the distance matrices, followed by eigen-decomposition and keeping the top d eigenvalues and eigenvectors to yield the embedding) followed by CCA (find two orthogonal $d \times d$ transformation matrices for each data set to maximize their correlation), which is a linear embedding and matching procedure.

Once the mappings are learned from the training data, the learned mappings ρ_l can be applied to match any new observations $y_1 \in \Xi_1$ and $y_2 \in \Xi_2$ of unknown correspondence, i.e., compute $\hat{y}_l = \rho_l(y_l) \in \mathbb{R}^d$, and declare y_1 and y_2 as matched if and only if $\hat{y}_1$ is sufficiently close to $\hat{y}_2$ in the common space. Ideally, a good matching procedure should be able to correctly identify the correspondence of the new observations, i.e., if the testing observations are truly matched in the context, the mapped points should be very close to each other in $\mathbb{R}^d$. If the testing observations are not matched, the mapped points should be far away from each other.

To evaluate a given matching algorithm, a natural criterion is the matching ratio used in seeded graph matching [27]. Assume that there exist multiple testing observations in each space; and for each testing observation y_1 in Ξ_1, there is a unique testing observation $y_2 \in \Xi_2$ such that $y_1 \sim y_2$. Then they are correctly matched if and only if $\hat{y}_2$ is the nearest neighbor of $\hat{y}_1$ among all other testing data from Ξ_2, and vice versa. The matching ratio equals the percentage of correct matchings, and a higher matching ratio indicates a better matching algorithm.

The matching ratio based on nearest neighbor is often conservative, and can be a very small number when matching disparate real datasets. In practice, it is often more interesting to consider all neighbors within a small threshold, or rank multiple neighbors up to a limit. To that end, the testing power of the statistical hypothesis H_0: $y_1 \sim y_2$ considered in [29] is another suitable criterion, which directly takes the Euclidean distance $\|\hat{y}_1 - \hat{y}_2\|$ as the test statistic. To estimate the testing power for given data, we first split all observations into matched training data pairs, matched testing data pairs, and unmatched testing data pairs. After learning ρ_l from the matched training data and applying them to all testing data, the test statistic under the null hypothesis can be estimated from the matched testing pairs, and the test statistic under the alternative hypothesis can be estimated from the unmatched testing pairs. The testing power at any type 1 error level is directly estimated from the empirical distributions of the test statistic, and a higher testing power indicates a better manifold matching algorithm.

We used both the testing power and the matching ratio for evaluation in the numerical experiments, and in most cases they yield the same interpretation regarding which method has a better matching performance. Note that if the critical value at a given type 1 error level is used as a distance threshold, the testing power equals the probability that the distance between the matched pair is no larger than the distance threshold. Since the matching ratio only considers the nearest neighbor of the matched pair, the testing power is never smaller than the matching ratio.

2.2. Main algorithms

Our methodology is henceforth referred to as MMSJ. Algorithm 1 serves to learn the matching transformations from the matched training data, while Algorithm 2 maps any testing observation onto the learned manifolds.

Given the distance matrices Δ_l for the training data $\{X_l, l = 1, 2\}$, we first construct an $n \times n$ binary graph G by k-nearest-neighbor using the sum of normalized distance matrices $\sum_{l=1}^{2} \frac{\Delta_l}{\|\Delta_l\|_F}$, i.e., $G(i, j) = 1$ if and only if $\sum_l \frac{\Delta_l(x_{il}, x_{jl})}{\|\Delta_l\|_F}$ is among the smallest k elements in the set $\{\sum_l \frac{\Delta_l(x_{il}, x_{ql})}{\|\Delta_l\|_F}, q = 1, \ldots, n\}$.

Next, for each modality X_l, we calculate the shortest-path distance matrix Δ_l^G based on the normalized Δ_l and the joint graph G, i.e., solve the shortest-path problem using the weighted graph $\frac{\Delta_l \circ G}{\|\Delta_l\|_F}$, where $\circ$ denotes the Hadamard product. Then we apply MDS to embed Δ_l^G into $\mathbb{R}^d$ for each l, followed by the Procrustes matching to yield the matched data $\hat{X}_l$, i.e., the Procrustes

[1] https://github.com/cshen6/MMSJ.