Pattern Recognition 73 (2018) 76–88

Contents lists available at ScienceDirect

Pattern Recognition

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

# Improved data visualisation through nonlinear dissimilarity modelling

Iain Rice

*Aston University, Aston Triangle, Birmingham, B4 7ET, UK*

ARTICLE INFO

*Article history:*
Received 17 March 2017
Revised 29 May 2017
Accepted 17 July 2017
Available online 2 August 2017

*Keywords:*
Dissimilarity
Visualisation
Multidimensional scaling
RBF network

ABSTRACT

Inherent to state-of-the-art dimension reduction algorithms is the assumption that global distances between observations are Euclidean, despite the potential for altogether non-Euclidean data manifolds. We demonstrate that a non-Euclidean manifold chart can be approximated by implementing a universal approximator over a dictionary of dissimilarity measures, building on recent developments in the field. This approach is transferable across domains such that observations can be vectors, distributions, graphs and time series for instance. Our novel dissimilarity learning method is illustrated with four standard visualisation datasets showing the benefits over the linear dissimilarity learning approach.

## 1. Introduction

Dimension reduction algorithms used to generate visualisations of high dimensional data require a chart of observations which must follow a global or local structure. The Sammon map [43], Stochastic Neighbour Embedding (SNE) [19] and variants, the Gaussian Process Latent Variable Model (GPLVM) [21], Generative Topographic Map (GTM) [7], Metric Multidimensional Scaling (MDS) and Curvilinear Component Analysis (CCA) [11] assume global Euclidean structure. Bregman divergences generate mappings with non-metric multidimensional scaling in [45–47] however the use of the Euclidean distance, as in standard MDS, remains. In the case where the observed data is known to sit upon a non-Euclidean manifold it is typically assumed that local regions of the manifold are Euclidean. Algorithms such as Locally Linear Embedding [42], Laplacian Eigenmaps [6], Riemannian Manifold Learning [26] and methods using geodesic distances based upon local Euclidean structure such as Isomap [49], the Geodesic Nonlinear Map [25] and Curvilinear Distance Analysis [24] rely on this property holding. Furthermore these algorithms require smooth continuity between local charts. This is known to not be the case where a manifold is, for instance, fractal or when observations are sparse and not true neighbours. As such the choice of local neighbourhood size parameters presents a challenge, causing the potential for short-circuits in neighbourhood graphs. The work of FINE [10] assumes that observations sit upon a statistical Riemannian manifold which is less restrictive than the Euclidean counterpart [3]. As such FINE uses an approximation to the Fisher Information metric to calculate local distances between observations, however each of the proposed approximations are not without limitations. In contrast the framework of [40] embeds non-Euclidean data onto a latent sphere of with calculated radius. This is in contrast to almost all other dimension reduction algorithms which do not restrict the structure of the latent space.

The latent variable models GTM and GPLVM assume that observations sit upon hyper-ellipses. In the GTM case this structure is treated as isotropic and as such suffers from the issues of hyper-spherical geometry (see [23] for details). The hyper-ellipse of GPLVM only permits dimensions between observations to be independent, a trait known to be false in many time series and image analysis domains for instance. These approaches are therefore incapable of constructing a reliable chart for complex datasets. In [23] it was demonstrated that dimension reduction algorithms relying on nonconvex optimisation of latent points, for instance MDS, CCA and GTM, perform superior to mappings using convex optimisation including PCA, LLE and Isomap.

An alternative approach is considered in [30,31] for the task of pattern discovery in large datasets. Local affinity patterns are identified across patterns and observed dimensions to convey significant attributes. The test for significance involves a Euclidean thresholding scheme over the cleaned graph weight matrix. The highlighted local affinities should be anomalies or sources of information, allowing the user to focus on a small subset of a large collection of data. In contrast the result of information visualisation is to utilise all attributes of observations and present the visual map over all datapoints to a human for interpretation. Such a weighting matrix as used in [30] can however be integrated within several visualisation frameworks when the weighting function is specified.

The notion and impact of non-Euclidean pattern analysis is discussed at length in [37]. Despite the fact that there are many

*E-mail address:* i.rice@aston.ac.uk

http://dx.doi.org/10.1016/j.patcog.2017.07.016

causes of non-Euclidean observations, frameworks to handle such datasets are still emerging and have not been widely adopted [12]. When the nature of an observed manifold is of unknown topology one naturally is unaware of the dissimilarity measure which charts the manifold. It is however possible to learn such a chart using a combination of a set of multiple dissimilarity measures, a dictionary. This is the approach of multiple kernel learning where kernels are combined linearly or nonlinearly in order to improve regression or classification performance (see [9,15] for an overview). Multiple kernel learning has also been implemented in the field of manifold learning [2]. Multi-feature kernels were developed in [55] to learn features for facial recognition based on a dictionary approach and discriminant analysis rather than dimension reduction. This notion is developed further in [1] where a sparse hierarchical dictionary based on Gaussian kernels is used for classification.

The tasks of regression and classification are by nature supervised. In this paper we consider the case of dimension reduction, in particular visualisation, which is unsupervised and as such mapping targets do not exist. The targets in this case are learned by minimisation of a mapping cost function such that neighbourhoods and the topological ordering of data is preserved. Non-Euclidean charts form the input to the visualisation framework of [27] relying on linear discriminant analysis. This linear approach does not generalise to nonlinear mappings. Another linear projection is used in [56] to map data whose dissimilarities are specified by a probabilistic measure. This approach cannot suitably map nonlinear structures, but the proposed measure can be incorporated into the framework of this paper. A distance metric learning approach is proposed in [51] focussing on clustering by adapting a kernel to learn a dissimilarity measure rather than fixed combinations of kernels. The clustering of data through spectral construction of kernels is detailed in [4] with links to Laplacian Eigenmaps, however this by nature learns a local descriptor of data. Using adaptive metrics [17] present an analogue of PCA with the goal of intrinsic dimensionality estimation rather than performing dimension reduction. The variables of interest in data are learned in [39] in a linear fashion prior to dimension reduction, however the relative significance of these features compared to one-another is not retained. Linear combinations of kernels form the basis of the nonlinear dimension reduction performed in [53] however the kernels used are restricted to polynomial functions and no learning of dissimilarities upon a manifold is performed. The Canonical Correlation Analysis approach of [57] linearly combines separate local and global kernels to perform dimension reduction which for certain kernel choices will behave like Isomap. A far more expressive linear combination of multiple kernels is presented in [33] where the weighting is fixed prior to dimension reduction. The work of [14] creates an ensemble of different clustering partitions, which may be non-Euclidean by nature, allows for more accurate clustering and classification.

In this paper we present a method for learning a chart based on a dictionary of dissimilarity measures whilst simultaneously constructing a nonlinear mapping. In [41] a linear combination of dissimilarity measures was used in this way and it was shown that the quality and interpretability of visualisations improved when the chart is learned. This paper builds on this linear model by learning a nonlinear combination of dissimilarity measures using a universal approximator. In order to show the improvements of this nonlinear learning of dissimilarities we use Elastic MDS as in [41] to provide a benchmark for our experimental results, however our approach generalises to other visualisation algorithms. In order to demonstrate the impact of our approach we generate visualisations of four standard datasets with Elastic MDS and Isomap. We assess the quality of our results with a visual comparison of the mapped latent variables, however as discussed in [52] quantitative comparison with visual quality metrics are not appropriate for non-Euclidean mappings.

## 2. The learning task

The aim of this paper is to accurately estimate the chart of an observed manifold without assuming a particular metric, but by learning a mixture from a fixed dictionary of dissimilarities. As a precursor we build on the work of [41] and therefore focus on the case of Elastic MDS [32] to perform a comparative analysis of our approach. We assess the performance of the constructed chart through visual analysis of an embedding of a dataset. This embedding need not be Euclidean in terms of Witney's embedding theorem [54] as visualisation would only be possible here if the intrinsic dimensionality of a dataset were 3-dimensional or less. Elastic MDS generates an embedding of a dataset, $X$, with $N$ observations by constructing a set of latent points, $Y \in \mathbb{R}^V$. As is typical for the task of visualisation we fix $V = 2$ in this paper, however our methods generalise trivially to other integer values for $V$.

A particular benefit of MDS methods is that $X$ need not be vectorial or even explicitly known, it is only required that the matrix of pairwise dissimilarities $D_x(i, j)$ between observations $X_i$ and $X_j$ is given. The latent points $\boldsymbol{y}_i$ corresponding to observation $X_i$ are learned through gradient descent of the Elastic MDS cost function:

$$E = \sum_{i,j<i} \frac{(D_x(i,j) - D_y(i,j))^2}{(D_x(i,j))^2}, \tag{1}$$

where $D_y(i, j)$ denotes the dissimilarity between the latent, visualised points $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. This measure is typically taken to be the Euclidean distance. Elastic MDS is distinct from the popular Sammon map due to the quadratic term in the denominator of Eq. (1), making the cost function more sensitive to local distances by stretching $D_x(i, j)$, hence the term elastic. This local focus naturally comes at the expense of global preservation.

For the case that $X$ consists of vectorial observations, $\boldsymbol{x}_i$, it is typically assumed that $D_x(i, j)$ is the Euclidean distance in the literature. This measure is only appropriate in the cases where the observed manifold is Euclidean. In the Riemannian or non-Riemannian manifold cases this distance function will give an incorrect approximation of distance. On statistical Riemannian manifolds the natural distance measure is known to be the Fisher Information Metric, which is typically approximated [10] using other divergence measures. The aim of this paper is to approximate the distance between observations:

$$D_x(i,j) = f\left(X_i, X_j\right). \tag{2}$$

In [41] the function $f$ is approximated using a linear combination of dissimilarity measures as a dictionary:

$$D_x(i,j) = \sum_{l=1}^{L} \alpha_l D^l(i,j), \tag{3}$$

where $\alpha_l$ is the weight corresponding to the $l$th dissimilarity measure, constrained such that $\alpha_l$ sums to unity. The dictionary of $L$ dissimilarity measures is user specified and the weights were learned during the optimisation of the Elastic MDS cost function in Eq. (1). These weights were optimised using gradient descent over Eq. (3) with respect to each factor $\alpha_l$ in order to find the optimal representation achieving a global minima. The dictionary-based approach is suited to situations where the natural metric of the observed data is unknown. In the regression or classification setting it would typically be assumed that the measure generating a chart over observations $X_i$ is that which achieves the highest predictive performance, however there is no guarantee that the measure which charts the manifold will be identified. For the