Neurocomputing 71 (2008) 2029–2036

Contents lists available at ScienceDirect

Neurocomputing

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

Isotree: Tree clustering via metric embedding

Bai Xiao [a,*], Andrea Torsello [b], Edwin R. Hancock [c]

[a] Department of Computer Science, University of Bath, Bath BA2 7AF, UK
[b] Dipartimento di Informatica, University "Ca' Foscari" of Venice, Via Torino 155, 30172 Venezia Mestre, Italy
[c] Department of Computer Science, University of York, York YO10 5DD, UK

ARTICLE INFO

Available online 4 March 2008

Keywords:
Graph clustering
Metric embedding
Spectral graph theory

ABSTRACT

One of the problems that hinders the spectral analysis of trees is that they have a strong tendency to be co-spectral. As a result, structurally distinct trees possess degenerate graph-spectra, and spectral methods can be reliably used to neither compute distances between trees nor to cluster trees. The aim of this paper is to describe a method that can be used to alleviate this problem. We use the ISOMAP algorithm to embed the trees in a Euclidean space using the pattern of shortest distances between nodes. From the arrangement of nodes in this space, we compute a weighted proximity matrix, and from the proximity matrix a Laplacian matrix is computed. By transforming the graphs in this way we lift the co-spectrality of the trees. The spectrum of the Laplacian matrix for the embedded graphs may be used for purposes of comparing trees and for clustering them. Experiments on sets of shock graphs reveal the utility of the method on real-world data.

1. Introduction

Tree structures have been used with considerable effect in computer vision to represent both object shape, scene structure and object articulation [27]. Examples include the use of shock trees to represent object boundary structure [27], the use of free-trees to represent human form [12] and the use of trees as compact image data-structures [34]. One of the problems that arises in the manipulation of large amounts of tree data is that of clustering. Although this task can be effected by applying pairwise clustering methods to the edit distance between trees, it does not allow the distribution of trees to be visualized or the effects of systematic changes in tree-structure to be assessed. Moreover, since computing the edit distance between trees relies on the availability of correspondences between nodes, and this is potentially an NP-hard problem, the computational overheads can be large.

One way to overcome the problem of computing the distance between discrete structures is to embed them in a low-dimensional space that minimizes the distortion. In this low-dimensional space, distances may be computed by taking a standard norm between the embedded pattern vectors. The problem of how to construct such an embedding has been the focus of activity in pattern recognition for several decades. For instance principal components analysis (PCA) projects pattern vectors into a low-dimensional space that maximally preserves the variance of the original data [13]. Multidimensional scaling (MDS), on the other hand, can be used to embed non-ordinal data into a low-dimensional space which preserves the relational pattern residing in the set of pairwise distances between data-items by minimizing the stress of the data [7]. However, these pattern analysis methods can only be applied for the data which is in vectorial form, or a distance function is to hand, and hence do not extend easily to discrete structures such as trees or graphs. In the mathematics literature, on the other hand, there is a considerable body of work aimed at understanding how graphs can be embedded in manifolds. Broadly speaking there are three ways in which the problem has been addressed. First, the graph can be interpolated by a surface whose genus is determined by the number of nodes, edges and faces of the graph. Second, the graph can be interpolated by a hyperbolic surface which has the same pattern of geodesic (internode) distances as the graph [1,5]. Third, a manifold can be constructed whose triangulation is the simplicial complex of the graph [33,21]. A review of methods for efficiently computing distance via embedding is presented in the recent paper of Hjaltason and Samet [11].

In the pattern analysis community, there has recently been renewed interest in the use of embedding methods motivated by graph theory. One of the best known of these is ISOMAP [30]. Here a neighborhood ball is used to convert data-points into a graph, and Dijkstra's algorithm is used to compute the shortest (geodesic) distances between nodes. The matrix of geodesic distances is used as input to MDS. The resulting algorithm has been demonstrated to locate well-formed manifolds for a number

* Corresponding author.
E-mail addresses: xb202@cs.bath.ac.uk (B. Xiao), torsello@dsi.unive.it (A. Torsello), erh@cs.york.ac.uk (E.R. Hancock).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.11.033

of complex data sets. Related algorithms include locally linear embedding [16] which is a variant of PCA that restricts the complexity of the input data using a nearest neighbor graph, and the Laplacian eigenmap that constructs an adjacency weight matrix for the data-points and projects the data onto the principal eigenvectors of the associated Laplacian matrix (the degree matrix minus the weight matrix) [2]. Collectively, these methods are sometimes referred to as manifold learning theories.

In this paper, we are interested in the problem of embedding trees in a pattern space for the purposes of both visualization and analysis (including clustering and classification). One of the methods that has proved effective for the embedding and pattern analysis of trees is spectral graph theory [6]. For instance, Dickinson and his co-workers [25,14] have shown how graph-spectra can be used to index shock-trees. There are two criticisms that can be leveled at the spectral analysis of trees. First, graphs that are not isomorphic can be co-spectral. As demonstrated by Schwenk et al. [23,3], due to their sparse edge-structure this problem is accentuated for trees. The second problem is the distortion produced by the embedding. In [14] the metric embedding algorithm gives a distortion that is proportional to $\sqrt{\log \log |A|}$, where A is a set of points in the original metric space, $|A|$ is the number of points in that space. In [9] the distortion is $\Omega(l(T)^{1/d})$, where $l(T)$ is the number of leaves in a tree T. So, when size of the trees becomes large, then the distortion will also become large.

To overcome these problems in this paper we investigate whether methods from manifold learning theory can be combined with spectral graph theory to develop effective tools for tree analysis. The idea is to use manifold learning methods to embed the trees in a low-dimensional space, and to perform spectral analysis on the co-ordinate data for the embedded tree-nodes. We proceed as follows. We commence by using a strategy similar to ISOMAP to embed the trees in a Euclidean pattern space. This is done by computing a matrix of shortest (geodesic) distances between nodes in the tree. We then apply MDS to the distance matrix, and this embeds the individual nodes of the tree in a Euclidean space. Once embedded in this space, we construct a weighted Laplacian matrix for the nodes of the tree by exponentiating the negative squared-distance between nodes. The spectrum of eigenvalues of the Laplacian can be used for the purposes of tree clustering and visualization.

2. Metric embedding of trees

The problem of embedding finite metric space into Euclidean spaces, or other normed spaces, that approximately preserve the metric is one that has received considerable attention in recent years. A number of ways have been proposed for measuring the quality of an embedding procedure. The *distortion* has been widely accepted as a measure of the quality of the embedding. For a finite metric space (X, d) and $c \geqslant 1$, there is an embedding φ of X into Y where for every two points $x_1, x_2 \in X$ satisfy the condition

$$d(x_1, x_2) \geqslant \|\varphi(x_1) - \varphi(x_2)\| \geqslant \frac{1}{c} d(x_1, x_2) \tag{1}$$

Such an embedding is said to be have *distortion* $\leqslant c$ [16]. Recently low-distortion embedding has provided powerful tools for designing efficient pattern analysis algorithms. This is because that they enable us to reduce problems defined over difficult metrics to problems over much simpler ones.

The starting point for most metric embedding methods is Bourgain's [4] Lemma:

> Any finite metric (X, d) can be embedded into l_2^p with $p < \infty$ with distortion $O(\log |X|)$.

We denote $\Re^n$ equipped with l_q norm by l_q^n. The Euclidean norm is l_2. The l_q norm is defined as $\|(x_1, \ldots, x_n)\|_q = (\sum |x_i|^q)^{1/q}$. The original bound on p proved by Bourgain was exponential with n and too large to be of practical use. We seek to introduce an embedding with a much lower distortion.

2.1. Metric embedding of trees by using isomap

We first define a suitable metric for the trees or graphs. For a given graph $G = (A, E)$, A represents the nodes in the graph and E represents the edge relations between the nodes. Suppose that D is a metric on the graph G. The metric must satisfy the condition that for any three vertices u, v and $w \in A$, if $D(u, v) = D(w, v) \geqslant 0$, then $D(u, u) = 0$ and $D(u, v) \leqslant D(u, w) + D(w, v)$. There are many ways to define metric distances on a graph. The best known is the shortest-path metric $D(u, v) = \delta(u, v)$, which is the shortest path distance between u and v for all $u, v \in A$. In fact, if the graph G is a tree, the shortest path between any two vertices is unique, and the weights of the shortest paths between any two vertices will define a metric $D(.,.)$. Since we can treat trees as a special kind of graph, we can use the shortest-path metric for trees.

Our goal is to find a low-distortion or distortion-free embedding from the tree metric space into a normed space. Here we use Isomap (isometric feature mapping) [30] as a way to solve the low-distortion tree embedding problem. The idea behind Isomap is to apply classical MDS [7] to map data points from their high-dimensional input space to low-dimensional coordinates of a nonlinear manifold. The key contribution is hence to apply MDS to the pairwise distances not in the input Euclidean space, but in the geodesic space of the manifold.

Although the method was originally devised for dimensionality reduction, we can use it here for the low-distortion tree embedding problem. Viewed as an isometric feature mapping, Isomap is a mapping $f : X \to Y$ from the observation space X to a Euclidean feature space Y that preserves as closely as possible the intrinsic metric structure of the observations, i.e. the distances between observations as measured along geodesic (shortest) paths of X [30]. The distortion c in this embedding is nearly 1.

For trees, the embedding procedure is straightforward. We first construct the shortest path distance matrix S for each tree. Each element d_{i_1,i_2} in S is the shortest path distance between the pair of nodes i_1 and i_2 of the tree. We embed each tree in a Euclidean space by performing MDS on the matrix S.

2.2. Multidimensional scaling

MDS is a procedure which allows data specified in terms of a matrix of pairwise distances to be embedded in a Euclidean space. The pairwise geodesic distances between nodes $d_{i1,i2}$ are used as the elements of an $N \times N$ dissimilarity matrix S, whose elements are defined as follows:

$$S_{i1,i2} = \begin{cases} d_{i1,i2} & \text{if } i_1 \neq i2 \\ 0 & \text{if } i_1 = i2 \end{cases} \tag{2}$$

In this paper, we use the classical MDS method. The first step of MDS is to calculate a matrix T whose element with row r and column c is given by $T_{rc} = -\frac{1}{2}[d_{rc}^2 - \hat{d}_{r.}^2 - \hat{d}_{.c}^2 + \hat{d}_{..}^2]$, where $\hat{d}_{r.} = (1/N)\sum_{c=1}^{N} d_{rc}$ is the average dissimilarity value over the rth row, $\hat{d}_{.c}$ is the similarly defined average value over the cth column and $\hat{d}_{..} = (1/N^2)\sum_{r=1}^{N}\sum_{c=1}^{N} d_{r,c}$ is the average similarity value over all rows and columns of the similarity matrix T.

We subject the matrix T to an eigenvector analysis to obtain a matrix of embedding co-ordinates X. If the rank of T is $k, k \leqslant N$, then we will have k non-zero eigenvalues. We arrange these k non-zero eigenvalues in descending order, i.e. $l_1 \geqslant l_2 \geqslant \cdots \geqslant l_k > 0$.