Eurographics Conference on Visualization (EuroVis) 2016
E. Bertini, N. Elmqvist, and T. Wischgoll
(Guest Editors)

Manifold Visualization via Short Walks

Y. Zhao[1], S. Tasoulis[2], and T. Roos[3]

[1]Department of Computer Science, Aalto University, Finland
[2]Department of Applied Mathematics, John Moores University, United Kingdom
[3]Department of Computer Science, University of Helsinki, Finland

Abstract

Visualizing low-dimensional non-linear manifolds underlying high-dimensional data is a challenging data analysis problem. Different manifold visualization methods can be characterized by the associated definitions of proximity between high-dimensional data points and score functions that lead to different low-dimensional embeddings, preserving different features in the data. The geodesic distance is a popular and well-justified metric. However, it is very hard to approximate reliably from finite samples especially between far apart points. In this paper, we propose a new method called Minimap. The basic idea is to approximate local geodesic distances by shortest paths along a neighborhood graph with an additional penalizing factor based on the number of steps in the path. Embedding the resulting metric by Sammon mapping further enhances the local structures at the expense of long distances that tend to be less reliable. Experiments on real-world benchmarks suggest that Minimap can robustly visualize manifold structures.

Categories and Subject Descriptors (according to ACM CCS): I.4.10 [Image Processing and Computer Vision]: Image Representation—Multidimensional

1. Introduction

Many high-dimensional data sets can be characterized by low-dimensional, locally linear but globally non-linear structures, i.e., manifolds. Examples include image, audio, and video data as well as many scientific data sets such as various sources of genomics and physics data. In many of these instances, conventional linear approaches such as principal component analysis (PCA), are insufficient due to non-linearities between the intrinsic structure of the manifold and the high-dimensional representation in the observed (ambient) space, see e.g. [LV07].

A large number of non-linear dimensionality reduction methods have been proposed. Among these methods, Isomap is a representive one. Instead of using the Euclidean distance, Isomap is based on approximating geodesic distance along the manifold. Examples of successful applications of Isomap include visualization of biomedical data [HILM09] and head pose estimation [RSH02]. However, Isomap is vulnerable to problems causes by "short-cuts" between points lying on different parts of the manifold [BST*02]. These problems are more severe for points that are far apart since the paths connecting them traverse through many intermediate points, each of which increases the inaccuracy.

Another successful approach is called neighborhood embedding (NE). It is based on a definition of proximity in terms of a probability measure defined by the pairwise Euclidean distances. The embedding is driven by a divergence measure that encourages preservation of local structure (short distances) at the expense of the global structure (long distances). A third idea, underlying a recent manifold clustering method [YHD*12], is to define proximities in terms of random walks instead of shortest paths. In order to retain locality, the random walks are adjusted by an attenuation factor that places more weight on walks with a small number of steps than on walks with many steps. This reduced the problems associated with Isomap when dealing with paths with a large number of steps.

In this paper, we try to combine the advantages of these three approaches. We argue that a combination of geodesic distance defined using shortest paths penalized by an attenuation factor that places more weight on paths with a small number of steps and a score function that leads to an embedding that focuses on the local structure, leads a robust manifold visualization. We demonstrate the proposed method, which we call Minimap (it emphasizes the small detail in the mapping), on common real-world benchmarks. The advantages of the method are seen both visually as well as in terms of a numerical score measuring separability.

The remainder of this paper is organizes as follows. We review relevant approaches related to our method in Sec. 2. In Sec. 3, we provide justification for the use of a limited number of steps of geodesic distance, and describe our Minimap method in detail. Experimental results are presented in Sec. 4.

DOI: 10.2312/eurovisshort.20161166

2. Preliminaries

We briefly present related previous methods more formally in order to introduce the building blocks of our method in the next section. Given a set of multivariate data points $\{x_1, x_2, ..., x_n\}$ in an ambient space R^M, we denote by D a matrix of pair-wise *proximities* (or similarities), so that the elements of the matrix d_{ij} represent the proximity between points x_i and x_j. The definition of proximity is an important element of the method. An embedding of the data onto a lower-dimensional space, R^m, is a mapping $x_i \mapsto y_i \in R^m$ for $i = 1, ..., n$ such that the similarities in the low dimensional space δ_{ij} approximate the proximities d_{ij}.

Dimension reduction techniques can be divided into several categories depending on one hand on whether the proximities are linear or non-linear in the distance between the points in the ambient space and on the other hand on whether they aim to preserve global or local structures [LV07].

MDS. Multidimensional Scaling (MDS) [CC00] is a linear, global structure preserving method. Both d_{ij} and δ_{ij} are commonly measured by Euclidean distance, $d_{ij} = \|x_i - x_j\|_2$, $\delta_{ij} = \|y_i - y_j\|_2$. The mapping is optimized by minimizing the score function $\sum_{ij}(d_{ij} - \delta_{ij})^2$. In practice, the mapped coordinates y_i can be obtained by finding eigenvectors of the similarity matrix using singular value decomposition.

Due to the squared distance used in the MDS score function, the importance of larger distances tends to be exaggerated compared to shorter distances. This can be rectified by the use of Sammon mapping [Sam69]. The idea is to assign each d_{ij} an associated weight w_{ij} that emphasizes small values of d_{ij}. The score function becomes $\sum_{ij} \frac{1}{w_{ij}}(d_{ij} - \delta_{ij})^2$. Usually the weight is given in the form $w_{ij} \propto d_{ij}^{-k}$ where $k = 1$ or $k = 2$. It has been shown that in many applications, Sammon mapping leads to improved local structure preservation, see e.g. [Ten98], [Ver06].

Isomap. Isomap [TDSL00] differs from MDS by using geodesic distance to compute pairwise input-space distances rather than simple Euclidean distances. To approximate the geodesic distances, a neighborhood graph is first constructed by connecting each point to its k nearest neighbours (in Euclidean distance). The geodesic distance is then obtained by computing the shortest path x_i to x_j along the neighborhood graph. The resulting geodesic distances are used to compute an embedding by MDS.

Yang [Yan04] also proposd a variant of Isomap where the MDS step is replaced by Sammon mapping in order to emphasize the local structure. However, we have observed that the variant is not competitive wrt. other existing techniques such as t-SNE (see below) or in fact even standard Sammon mapping in the benchmarks we use.

t-SNE. t-SNE [VdMH08] is a recent nonlinear, local structure preserving method that has become a popular visualization technique. t-SNE preserves similarities between points using a probabilistic formulation. It first computes a conditional probability distribution using a Gaussian kernel as follows:

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

where p_{ij} stands for the probability of point x_i to be a neighbour of point x_j in the high dimensional space. In the desired low dimensional space, the same is done but the kernel is changed to Student's t-kernel.

$$q_{i,j} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

t-SNE measures the distance between these two distributions by the Kullback-Leibler (KL) divergence. Due to the properties of the KL-divergence, t-SNE pays more attention to short distances than methods such as MDS. This is further emphasized by the heavy tail property of the t-distribution.

3. The Minimap method

Let us first motivate the choices we make in designing a new dimension reduction and visualization method. The main goal of the method is to quickly gain understanding of the coarse structure of a large, high-dimensional dataset. We particularly focus on preserving the identity of locally compact subgroups, or clusters, that may exist in the data, even though we do not assume this to be the case.

3.1. Proximity by random walk

To illustrate different proximity measures, L. Yang [Yan04] and Z. Yang *et al.* [YHD*12] consider an idealized situation where the neighborhood graph should be a block-diagonal matrix as shown in Figure 1(a). In the matrix, rows and columns corresponding to instances in the same cluster should have large values whereas entries corresponding to instances in different clusters should have low values.

If the similarities are defined directly in terms of a k-nearest neighbor graph, the similarities between non-neighboring pairs are set to zero. Figure 1(b) shows an example of a neighborhood graph built by symmetrical 5-nearest neighbors. As we can see, compared to the ideal case, this matrix is very sparse, i.e., not all pairs of instances in the same cluster are connected.

An alternative way to measure proximity is to use random walks as proposed by [YHD*12]. Denote $Q = D^{-1/2}SD^{-1/2}$ the normalized similarity matrix, where D is diagonal matrix with $D_{ii} = \sum_j S_{ij}$. We can then assign a decaying weight $\alpha \in (0,1)$ to each step so that each subsequent step has less effect on the end result. In each step, the proximity is increased by $(\alpha Q)^j$, where j represents the j^{th} step. In the limit $j \to \infty$, we get $S' = \sum_{j=0}^{\infty}(\alpha Q)^j = (I - \alpha Q)^{-1}$. To compute S', a matrix inverse operation is required which is of complexity $O(n^3)$, which may be impractical for large matrices.

3.2. Proximity by directed walk

Inspired by the random walk idea, we propose a novel variant of local geodesic distance as the proximity measure. First, to avoid the accumulation of approximation error along long geodesic paths, we adapt the idea of an attenuation factor on each step of the walk from

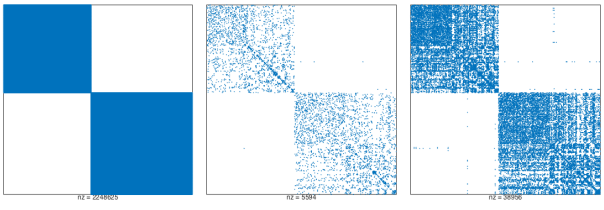

Figure 1: Illustration of similarity graphs between handwritten digits 1 & 2 from MNIST dataset: *Left:* the correct clusters in the data, *Middle:* symmetrized 5-NN graph, *Right:* constructed graph after three steps of directed walk.

Z. Yang *et al.*. However, instead of considering *random* walks originating from point x_i, many of which are irrelevant when the target is point x_j, we consider a walk directed toward x_j. The selected path is the one that has the least weighted distance where the edge lengths are magnified as a function of the number of steps on the path. This can reduce the approximation error by focusing on paths with fewer steps.

In practice, a simple and robust heuristic is to assign equal proximity to all paths with no more than L steps and to ignore all paths with more than L steps. This way, the proximity matrix only depends on local connectivity within a radius of L steps. The constructed proximity matrix with $L = 3$, Fig. 1(c), is much denser than the 5-nearest neighbor graph. Another key benefit compared to computing the full geodesic distance, including paths with any number of steps, is computational efficiency.

Computing all pairwise geodesic distance takes $O(kN^2 \log N)$ by using Dijkstra's algorithm, while local geodesic distances up to paths with L steps can be implemented using only $O(k^L N)$. For each point, we store its k nearest neighbors in a dictionary structure. Then we expand a point's dictionary using depth-limited breadth-first search. After L expansions, we can get all neighbours that can be connected with this point. Assuming that dictionary look-up and insertion operations can be implemented in $O(1)$ time leads to the claimed time complexity $O(k^L N)$.

3.3. The Minimap procedure

We are now ready to piece things together to complete the Minimap method. The method consists of three stages. First, the proximity matrix D is populated by letting the proximity between points reachable by short directed walks be equal to a constant λ, as described in the previous subsection. Second, the remaining proximities between points for which a short walk doesn't exist, are defined in order to complete the matrix. Here we use another constant greater than λ. The ratio between the two constants turns out to be a critical choice. We have observed that no fixed constant independent of the size of the problem is suitable in all situations. The reason for this is that as the size of the data set grows, the number of unconnected points grows much faster than the number of connected points, and the importance of the connected points vanishes unless their relative weight is increased accordingly. Hence, we propose the default choice of $\delta_{ij} = \lambda = \log_{10}^2(n)/n$, where n is the number of data points, for points connected by a short walk, and the choice $\delta_{ij} = 1$ for points unconnected by a short walk.

The third and final stage of Minimap is the embedding. For the reasons outlined above, we adopt Sammon mapping as the embedding method. Note that letting λ decrease as the sample size grows implies that the weight $w_{ij} = d_{ij}^{-k}$ assigned to the distances between connected points increases with the sample size. Another heuristic choice that we found out to be useful is to first run MDS for 10 iterations with weights given by w_{ij}^{-2} to get the rough structure initialized, and then to switch to weights w_{ij}^{-1} until convergence. As with the t-SNE method, the embedding stage is the computational bottleneck.

4. Experiments

Datasets. Here we show experimental results on three datasets: i) The `Coil-20` dataset is a collection of grayscale images of 20 different objects. For each object, 72 pictures were taken in different orientations. The objects are uniformly scaled to fit within a 128×128 bounding box. ii) The `Umist` face dataset is another collection of grayscale images, consisting of a total of 575 images of 20 different persons, each represented as a 112×92 pixel image. iii) The `USPS` digits dataset contains images of handwritten digits. Each image is of size 16×16. Here we used a subset containing the digits 2, 4, 6, 8. All datasets are used directly without any preprocessing.

Experiment setup. In the experiments, we compute results using Isomap, t-SNE, and our Minimap method. For Isomap and t-SNE we use the default settings. For Minimap we limit the short walks to $L = 4$ steps and $k = 7$ for defining the k-nearest neighbor graph. (Due to restricted space, we omit a sensitivity analysis showing that these choices are not critical.) The other settings were as described above.

Figure 2 presents the mapping results of different methods on the three datasets. To measure the projection results numerically, we the use the standard adopted by [EV11], i.e., the percentage of the $k = 5$ nearest neighbours from the same class of points (defined in each data set as the set of points corresponding to a single object, person, or digit).

In both the `Coil-20` (top) and `Umist` (middle) data sets, the visualized points represent images of objects shown from different angles, and hence, it is natural that they form either circular or linear structures parametrized by the angle. In the `Coil-20` data, the t-SNE method captures some of the circular structures but misses most of them. In the `Umist` data, some of the linear structures, each of which is shown in a single color, are broken up by t-SNE. In both cases, Isomap provides an inferior result, but Minimap performs well. In the `USPS` data, none of the methods are able to extract further structure than the differences between the clusters corresponding to different digits.

Acknowledgments

This work was carried out while all authors were at the Department of Computer Science, University of Helsinki. This work was supported by the Academy of Finland (COIN CoE).

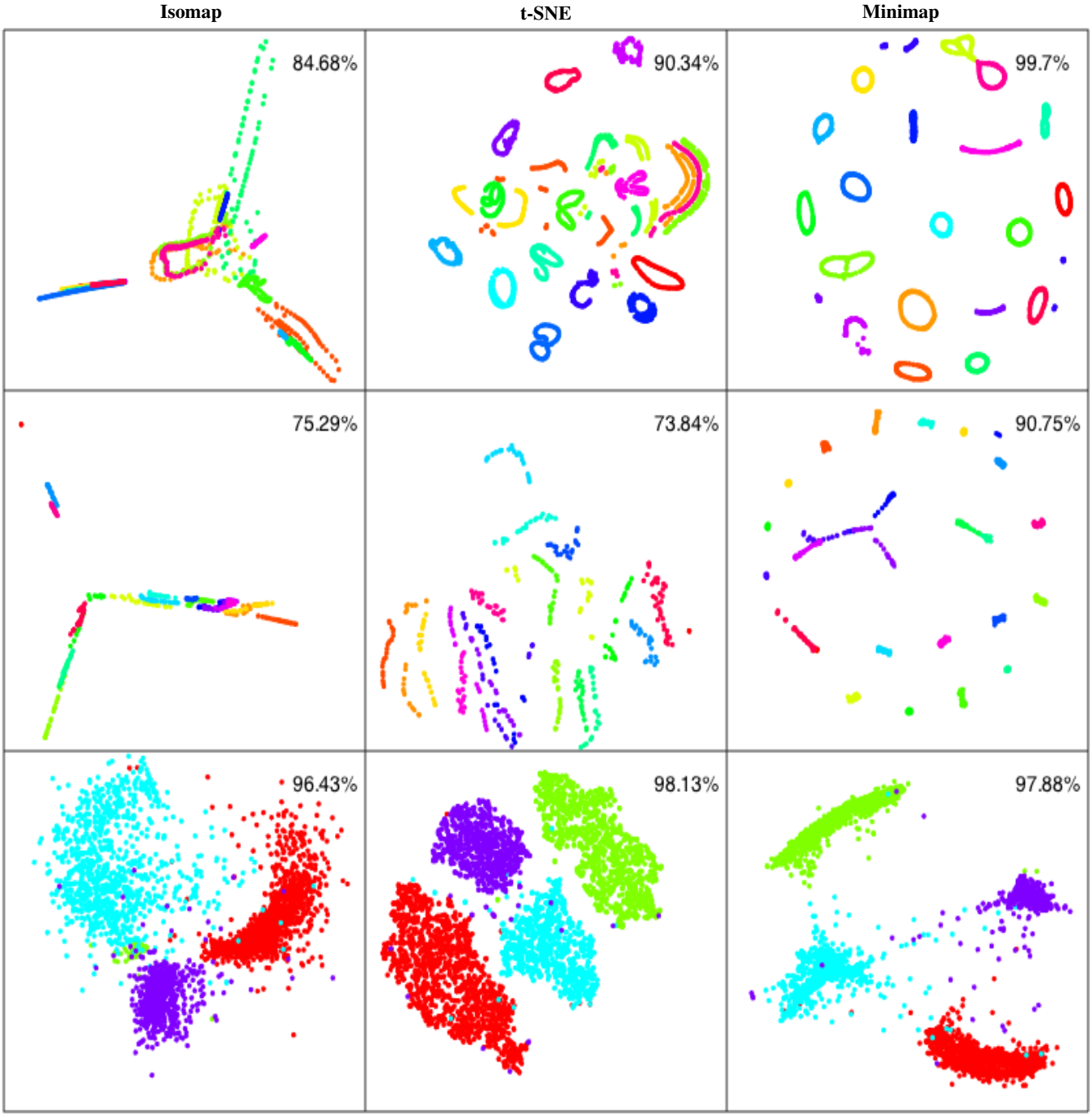

Figure 2: Visualizations of three data sets (from top to bottom) `Coil-20`, `Umist`, `USPS` obtained by Isomap (left column), t-SNE (middle), and our Minimap method (right). Ground truth is shown in different colors. Accuracy is showed at the top right corner of each panel.

References

[BST*02] BALASUBRAMANIAN M., SCHWARTZ E. L., TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: The Isomap algorithm and topological stability. *Science 295*, 5552 (2002), 7–7. 1

[CC00] COX T. F., COX M. A.: *Multidimensional Scaling*. CRC Press, 2000. 2

[EV11] ELHAMIFAR E., VIDAL R.: Sparse manifold clustering and embedding. In *Advances in neural information processing systems* (2011), pp. 55–63. 3

[HILM09] HAAS P. J., ILYAS I. F., LOHMAN G. M., MARKL V.: Discovering and exploiting statistical properties for query optimization in relational databases: A survey. *Statistical Analysis and Data Mining: The ASA Data Science Journal 1*, 4 (2009), 223–250. 1

[LV07] LEE J. A., VERLEYSEN M.: *Nonlinear Dimensionality Reduction*. Springer Science & Business Media, 2007. 1, 2

[RSH02] ROWEIS S. T., SAUL L. K., HINTON G. E.: Global coordination of local linear models. *Advances in Neural Information Processing Systems 14* (2002), 889–896. 1

[Sam69] SAMMON J. W.: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*, 5 (1969), 401–409. 2

[TDSL00] TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (2000), 2319–2323. 2

[Ten98] TENENBAUM J. B.: Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems* (1998), 682–688. 2

[VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 2579-2605 (2008), 85. 2

[Ver06] VERBEEK J.: Learning nonlinear image manifolds by global alignment of local linear models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 8 (2006), 1236–1250. 2

[Yan04] YANG L.: Sammon's nonlinear mapping using geodesic distances. In *ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition* (2004), vol. 2, IEEE, pp. 303–306. 2

[YHD*12] YANG Z., HAO T., DIKMEN O., CHEN X., OJA E.: Clustering by nonnegative matrix factorization using graph random walk. In *Advances in Neural Information Processing Systems* (2012), pp. 1079–1087. 1, 2