Mercer Kernel Isomap

최희열[O], 최승진
포항공과대학교 컴퓨터공학과
{hychoi[O], seungjin}@postech.ac.kr

Mercer Kernel Isomap

Heeyoul Choi[O], Seungjin Choi
Department of Computer Science, POSTECH, Korea

Abstract

Isomap [1] is a manifold learning algorithm, which extends classical multidimensional scaling (MDS) by considering approximate geodesic distance instead of Euclidean distance. The approximate geodesic distance matrix can be interpreted as a kernel matrix, which implies that Isomap can be solved by a kernel eigenvalue problem. However, the geodesic distance kernel matrix is not guaranteed to be positive semidefinite. In this paper we employ a *constant-adding* method, which leads to the Mercer kernel-based Isomap algorithm. Numerical experimental results with noisy "Swiss roll" data, confirm the validity and high performance of our *kernel Isomap* algorithm.

1. Introduction

Manifold learning involves inducing a smooth nonlinear low-dimensional manifold from a set of data points drawn from the manifold. Isomap is a representative isometric mapping, which extends metric MDS by considering approximate geodesic distance, instead of Euclidean distance [1].

Classical scaling (that is one of metric MDS) is closely related to principal component analysis [2]. The projection of the centered data onto the eigenvectors of the data sample covariance matrix, returns the classical solution. Classical scaling with Euclidean distances as the dissimilarities, is explained in the context of PCA, so that it provides a generalization property (or projection property) where new data points (which do not belong to a set of training data points) can be embedded in a low-dimensional space, through a mapping computed by PCA. In the same manner, a non-Euclidean dissimilarity can be used, although there is no guarantee that the eigenvalues are nonnegative. A relationship between kernel PCA and metric MDS were investigated in [3].

The approximate geodesic distance matrix used in Isomap, can be interpreted as a kernel matrix [4]. However, the kernel matrix based on the doubly centered approximate geodesic distance matrix, is not always positive semidefinite. We exploit a constant-adding method such that the geodesic distance-based kernel matrix is guaranteed to be positive semidefinite. Mercer kernel-based Isomap algorithm has a generalization property so that test data points can be successfully projected using a kernel trick as in kernel PCA [5], whereas in general embedding methods (including Isomap) do not have such a property. Also, in this paper, we compare several methods to make kernel matrix to be positive semidefinite.

2. Kernel Isomap

2.1. Isomap as Kernel Method

As in multidimensional scaling (MDS), Isomap first constructs a matrix of pairwise distances between the different data points. However, instead of directly using Euclidean distance in the high-dimensional space, Isomap uses an approximation of geodesic distance. First, it constructs a symmetric adjacency graph using criteria such as symmetric nearest neighborhoods or ε-ball neighborhoods based on Euclidean distance. Then Dijkstra's algorithm is used to compute the shortest path among edges in the neighborhood graph to define the total distance between pairs of points. Finally, MDS is applied to this shortest path distance matrix.

As pointed out in [3], metric MDS can be interpreted as kernel PCA. In a similar fashion, Isomap can be considered as a kind of kernel method [4]. We can take the approximated distances D used in Isomap and consider the following kernel:

$$K = -\frac{1}{2} H D^2 H, \tag{1}$$

where D^2 means element-wise square of D, H is the centering matrix, given by $H=I-ee^T/N$ and $e = [1,1,\cdots,1]^T \in R^N$.

However, this kernel is not guaranteed to be positive semidefinite. The reason why the kernel matrix of Isomap is not positive definite in the smooth manifold, is mainly the approximation of the geodesic distance and noise. So, we propose kernel Isomap which has the noise robustness and projection property.

2.2. Kernel Isomap Algorithm

Given N objects with each object being represented by an m-dimensional vector x_i, $i=1,\cdots,N$, the kernel Isomap algorithm finds a mapping which places N points in a low-dimensional space. In manifold learning, it is assumed that the manifold is generated by embedded low-dimensional data by some mapping. Two theorems stated below show the necessity and possibility of the positive semidefinite kernel matrix.

Theorem 1. [6] Given an arbitrary map ϕ into a feature space H, the matrix K,

$$K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle, \tag{2}$$

is positive semidefinite.

Theorem 1 means that if we calculate the kernel from the manifold learning, then that kernel has to be positive semidefinite. Now, with the kernel of Isomap which is not positive semidefinite, we have to find the positive semidefinite kernel to reflect the manifold properly. The positive semidefinite property guarantees the embedded manifold according to the kernel.

Theorem 2. [7] Given an arbitrary (possibly non-metric) (N x N) dissimilarity matrix D with zero self-dissimilarities, there exists a transformed matrix $\tilde{D}$ such that the matrix $\tilde{D}$ can be interpreted as a matrix of Euclidian distances between a set of vectors $\{x_1, x_2, \cdots, x_N\}$. $\tilde{D}$ is derived from D by both symmetrizing and applying the constant shift embedding trick.

In Theorem 2, the modified dissimilarity, $\tilde{D}$, guarantees the positive semidefinite kernel and low-dimensional embedded manifold. To calculate the $\tilde{D}$ stated in Theorem 2, an analytic solution was proposed [8][2]. The algorithm is below:

Algorithm Outline: Kernel Isomap

Step 1. Identify the k nearest neighbors (or ε-ball neighborhood) of each input data point and construct a neighborhood graph where edge lengths between points in a neighborhood are set as their Euclidean distances.

Step 2. (Shortest Path Problem) Compute approximate geodesic distances, d_{ij}, containing shortest paths between all pairs of points.

Step 3. Construct a kernel matrix $K(D^2)$ based on the approximate geodesic distance matrix D^2 as Eq. (1).

Step 4. Compute the largest eigenvalue, c^*, of the matrix

$$\begin{bmatrix} 0 & 2K(D^2) \\ -I & -4K(D) \end{bmatrix}, \tag{3}$$

and construct a Mercer kernel matrix $\tilde{K} = \tilde{K}(\tilde{D}^2)$ by

$$\tilde{K} = K(\tilde{D}^2) + 2cK(D) + \tfrac{1}{2}c^2 H, \tag{4}$$

where $\tilde{K}$ is guaranteed to be positive semidefinite for $c \geq c^*$.

Step 5. Compute the top n eigenvectors of $\tilde{K}$, which leads to the eigenvector matrix $V \in R^{N \times n}$ and the eigenvalue matrix $\Lambda \in R^{n \times n}$.

Step 6. The coordinates of the N points in the n-dimensional Euclidean space are given by $Y = \Lambda^{1/2} V^T$.

A main difference between the conventional Isomap and our kernel Isomap, lies in Step 4 which is related to the *additive constant problem* that was well studied in metric MDS. The additive constant problem aims at finding a value of constant, c, such that the dissimilarities defined by

$$\tilde{d}_{ij} = d_{ij} + c(1 - \delta_{ij}), \tag{5}$$

have a Euclidean representation for all $c \geq c^*$ and δ_{ij} is the Kronecker delta. Substituting $\tilde{d}_{ij}$ for d_{ij} in Eq. (5) gives Eq. (4). For $\tilde{K}$ to be positive semidefinite, it is required that $x^T \tilde{K} x \geq 0$ for all x. Cailliez showed that c^* is given by the largest eigenvalue of the matrix (3) (see Sec. 2.2.8 in [2]).

The matrix $\tilde{K}$ is a Mercer kernel matrix, so its (i,j)-element is represented by

$$\tilde{K}_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, \tag{6}$$

where ϕ is a nonlinear mapping onto a feature space or a low-dimensional manifold. The coordinates in the feature space can be easily computed by projecting the centered data matrix onto the normalized eigenvectors of the sample covariance matrix in the feature space,

$$C = \tfrac{1}{N}(\Phi H)(\Phi H)^T, \tag{7}$$

where $\Phi = [\phi(x_1), \ldots, \phi(x_N)]$.

2.3. Generalization Property

Isomap is one of kernel methods like kernel PCA. But, the main difference is that kernel PCA also provides an embedding for test points, whereas Isomap only embeds the training points. As kernel PCA, the kernel Isomap provides a *generalization property (projection property)*, which provides a mapping for a test data point. If new test data points $t_1, t_2, \cdots, t_L$ are given, we can project the test data onto the low-dimensional embedded manifold. First, we need to calculate the kernel matrix of test data. So, geodesic distance d_{Tij} from test data point t_i to the training data x_j, $j=1,\ldots,N$, is obtained. Then the following relation in the data space is derived,

$$d_{Tij}^2 = \langle t_i, t_i \rangle + \langle x_j, x_j \rangle - 2\langle t_i, x_j \rangle \cdot \quad (8)$$

We can derive the test kernel matrix, $K_T \in R^{L \times N}$, using the kernel matrix of training data in Eq. (4)

$$\langle t_i, x_j \rangle_\phi = -\frac{1}{2}\left(d_{Tij}^2 - \frac{1}{N}\sum_{k=1}^{N} d_{Tik}^2 + \frac{1}{N}\sum_{l=1}^{N} \tilde{K}_{ll} - \tilde{K}_{jj} \right), \quad (9)$$

where $\langle t_i, x_j \rangle_\phi$ is the (i,j)-element of the test kernel matrix, K_T. With K_T, we can derive the centered kernel matrix as in kernel PCA,

$$\tilde{K}_T = K_T - e_L \tilde{K} - K_T e_N + e_L \tilde{K} e_N, \quad (10)$$

where all entries of $e_L \in R^{L \times N}$ and $e_N \in R^{N \times N}$ are $1/N$. Here, $\sum_{j=1}^{N} \tilde{K}_{Tij} = 0$, since K_T is the kernel matrix of centered data. Finally, the projection of test data are

$$y_l^j = \sum_{i=1}^{N} V_{ij} \tilde{K}_{Tjt}, \quad (11)$$

where V_{ij} is the ith-element of V_j in Step 5 and y_l^j is the jth-element of y_l.

3. Numerical Experiments

We compared our kernel Isomap algorithm to the conventional Isomap algorithm, using Swiss roll data that was also used in Isomap. Noisy Swiss roll data was generated by adding isotropic Gaussian noise with zero mean and 0.25 variance (see Fig. 1 (a)). In the training phase, 1200 data points were used and the neighborhood graph was constructed using $k=4$ nearest neighbors of each data point, respectively. As in Isomap, the shortest paths were computed using the Dijkstra's algorithm, in order to calculate approximate geodesic distances.

An exemplary embedding result (onto 3-dimensional feature space) for Isomap and kernel Isomap, is shown in Fig. 1 (b) and (c). Though the results were not presented in this paper, for the case of noise-free Swiss roll data and noisy semi-sphere data, our kernel Isomap algorithm outperformed Isomap. The generalization property of our kernel Isomap is shown in Fig. 1 (d) where 3000 test data points are embedded with preserving local isometry well. In this figure, comparing (c) with (b), we can also see the noise robustness of kernel Isomap. Even though the conventional Isomap also looks like robust algorithm in 2 dimensional embedded manifold, in 3 dimensional space, it is not robust any longer. The manifold in (b) is not smooth while (c) is smooth.

4. Conclusion

We have presented the kernel Isomap algorithm where the approximate geodesic distance matrix was interpreted as a kernel matrix and an adding-constant method was exploited so that the geodesic distance-based kernel became Mercer kernel. Also, we compared other methods with adding-constant method to make the kernel matrix positive semidefinite. Main advantages of the kernel Isomap could be summarized as follows: (a) generalization property (i.e., test data points can be projected onto the feature space using the kernel trick as in kernel PCA); (b) robustness for low-level noisy data. The generalization property will derive the kernel Isomap to be useful for pattern recognition problems.

Acknowledgment

This work was supported by Brain Neuroinformatics program, Systems Bio-Dynamics Research Center under NCRC program, KOSEF 2000-2-20500-009-5 and Korea Ministry of Science and Technology under International Cooperative Research Program.

References

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, pp. 2319-2323, 2000.

[2] T. Cox and M. Cox, Multidimensional Scaling, 2nd ed. Chapman & Hall, 2001.

[3] C. K. I. Williams, "On a connection between kernel PCA and metric multidimensional scaling," Machine Learning, vol. 46, pp. 11-19, 2002.

[4] J. Ham, D. D. Lee, S. Mika, and B. Sch¨olkopf, "A kernel view of the dimensionality reduction of manifolds," in Proc. Int'l Conf. Machine Learning, Banff, Canada, 2004, pp. 369-376.

[5] B. Sch¨olkopf, A. J. Smola, and K. R. M¨uller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, no. 5, pp. 1299-1319, 1998.

[6] B. Sch¨olkopf and A. J. Smola, Learning with Kernels. MIT Press, 2002.

[7] V. Roth, J. Laub, J. M. Buhmann, and K. R. M¨uller, "Going metric: Denoising pairwise data," in Advances in Neural Information Processing Systems, vol. 15. MIT Press, 2003.

[8] F. Cailliez, "The analytical solution of the additive constant problem," Psychometrika, vol. 48, no. 2, pp. 305-308, 1983.

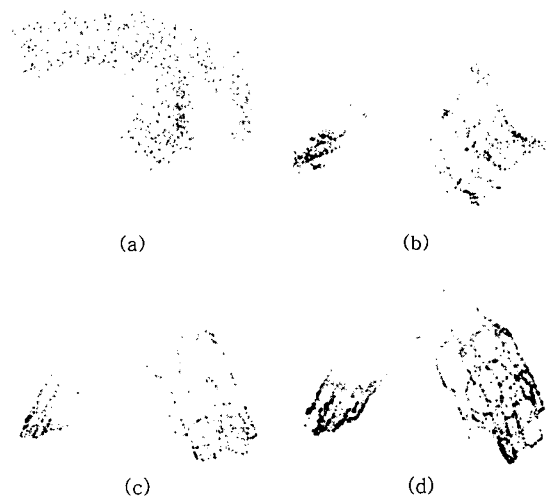

Fig. 1. Comparison of the conventional Isomap with our kernel Isomap for the case of *noisy Swiss Roll* data: (a) noisy Swiss Roll data; (b) embedded points using the conventional Isomap; (c) embedded points using our kernel Isomap; (d) projection of test data points using the kernel Isomap.