The Sea Surface Temperature Classification based on ISOMAP Analysis

John Chien-Han Tseng

Central Weather Bureau

Abstract

ISOMAP (isometric feature mapping) is one kind of nonlinear dimensionality reduction techniques for extracting the features in spatial-temporal data. Traditional linear dimensionality reduction like principal component analysis (PCA) measures the data by the Euclidean distance, which somehow cannot reflect the real structure in the given data. In brief, the PCA uses the linear framework to describe the data. The ISOMAP measures the data based on the geodesic distance, which gets the better real distance by the view of tracing along the manifold. The traditional PCA based on the sea surface temperature (SST) data with the Niño 3.4 index can be one of the classification tools to differentiate the ENSO events and the non-ENSO events. However, the classification results always cannot represent the differences (the physical distances) well or cannot distinguish the events effectively. That is, to understand the differences between the ENSO events, the La Niña events, or the normal events cannot only rely on the traditional PCA. In this report, we use the ISOMAP to classify the different SST monthly patterns. The ENSO events, the normal events, and the La Niña events reconstructed by the ISOMAP are easier to classify or to cluster than the constructed structures by the traditional PCA. At last, the eigenmodes (empirical orthogonal function, EOF) comparisons between the ISOMAP and the traditional PCA are shown in this report.

摘要

ISOMAP 等距特徵映射是一種非線性降維度的技術，用來擷取時間空間資料的特徵。傳統上線性降維的方法如主成分分析等，所量度資料間的關係是架構在歐式距離（一種平面）上，某種程度上不能反應資料關係可能是非線性、非平面距離的關係。ISOMAP 量度的距離是所謂的測地線距離(geodesic distance) ，也就是在局部小範圍的地區是採用平面的距離關係，但是整體上會跟著流形(manifold)變化計算資料兩點的距離。傳統上主成分分析，配合海溫資料與 Niño 3.4 指標值可以用來分類 ENSO 事件或是非 ENSO 的事件。但是，這種分類的方法常常不能將不同的事件差異有效分別好。也就是說，想要瞭解 ENSO 事件，La Niña 事件，或是正常年事件的彼此差異，不能完全依賴傳統的主成分分析。在這個報告中，我們使用 ISOMAP，透果每個月的海溫所重構的結構，比傳統主成分分析，更能凸顯 ENSO，La Niña 和正常年事件的差異。最後，ISOMAP 和傳統主成分分析的特徵模（經驗正交函數）異同之處也會於報告中說明。

1. Introduction

The goal of the classification or the clustering is to understand the differences among the events, e.g., the numerical data, the colors, the object shapes or figures, and etc. Now the question is: how to describe the difference, or how to measure the difference, and how to recognize the difference effectively. Traditionally, the analyzed data will be dimensional reduction to low dimension 2D or 3D points and the difference can be measured by the distance in the low dimensional space. Next, the key point is how to get the meaningful low dimensional space.

One of the traditional dimensionality reductions is the principal component analysis (PCA). The leading PCA components extract the main variances of the original data. The variances extracted by the leading PCA components are called explained variances. The fewer leading PCA components and the more explained variances, the better results we can get from the PCA. For example, if the original data dimensions are 10,000 and then we can just use three leading PCA components to explain the 80% original variances, the PCA results are very good. On the contrary, the worse situation is that the excess 100 leading PCA components to explain just 50% the original variance. Inevitably, this situation always happens in the real data analysis. So, there are many modifications of the PCA analysis (Alpaydin, 2010; Hsieh, 2009; Bishop, 2006). Once one gets the pretty good low dimensional PCA components, he/she cannot guarantee to get the well classification results. There low dimensional points from PCA can be classified well or not depend on the spread of the points enough or not.

In order to solve the classification problem, or for getting the well spread of the low dimensional points, Tenebaum *et al.* (2000) propose the isometric mapping (ISOMAP) to get the well spread of the low dimensional data points. They point to the traditional PCA taking the data linearly; for example, the time evolution is resolved by the linear evolution of the original data arrangement. We can image that the geopotential height evolves in one month by daily data. The 30 times data we taking is constrained by the linear time variation. We know the geopotential height will not evolve linearly in one month. But when we use the PCA, the covariance matrix of geopotential height is counted by linear consideration. When we use the linear coordinates to check the nonlinear variation, we will find the data points will concentrate in some places and separate in some places. The concentrating data points are not easy to classify and cause the classification fail. The ISOMAP tries to build the original nonlinear relations in the data by establishing the nearest neighbors. The ISOMAP keeps the small domains (manifold) linearity but reflect the larger domain nonlinear variations.

In this report, we use the sea surface temperature (SST) data to do the traditional ENSO classification. The comparisons of classification results by traditional PCA and the ISOMAP are shown.

2. Data and Methods

The SST data we used are the version 5 of NOAA NCDC ERSST (Extended Reconstructed global Sea Surface Temperature data based on COADS data). The time is from Jan. 1980 to

Apr. 2021. The ENSO, the normal, and the La Niña events are based on NOAA's climate prediction center Niño 3.4 index. The moving three months average SST excess 0.5 was determined to be the ENSO events. The ENSO events will be marked by red color, the normal events will be marked by yellow color, and the La Niña events will be marked by blue color. The different color represents the different label which will be used in the later classification.

The concept of ISOMAP is shown in Fig. 1.

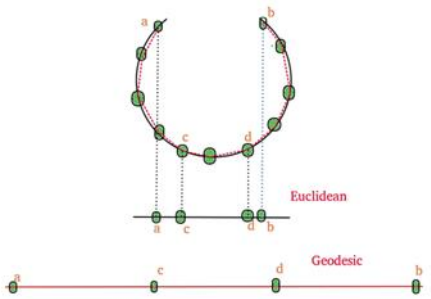

Fig. 1: The concept of the ISOMAP.

The real data points are located in the warp surface, which are shown in the arc curve in the Fig.1. When we take the PCA calculation, we assume the variation is linear, the relation between data points is like the short straight line. The PCA relation can be regarded as one kind of Euclidean distance. But the real distance between the point a and the point c is larger than the Euclidean distance. The PCA always fails to show the real situation. If we need to use the linear tool like linear algebra eigen solutions, we need to rearrange the relation, the distance, to the longer straight line in Fig. 1. The distance between the point a and the point b on the geodesic line truly reflects the distance in the warp surface, the curve, in the Fig. 1.

Tenebaum *et al*. (2000) propose to build the nearest neighbor graph, which is used to reflect the distance on the warp surface. That means we does not count the distance between the point a and c directly. The distance between the point a and the point c must pass by the other three points. There is no 'shortcut' between the point a and the point c. The geodesic distance is calculated on this neighbor graph and then we use this distance relation and follow the traditional PCA procedure to solve the eigen problem.

The classifier we used in this report is SSVM, one kind of the support vector machines (Lee and Mangasarian, 2000). All the testing results are from 20 times 5-fold cross validation.

3. Classification

We chose the leading three eigen components from the PCA and the ISOMAP to show the dimensionality reduction, and we will use the leading 20 eigen components to test the classification results. The PCA three leading eigen components were shown in Fig. 2. The Fig. 2a showed the 3D structure, and the Fig. 2b showed the 2D structure. In Fig. 2 we can say that most ENSO, normal, and La Niña events were already separated well. That means Meteorologists using the Niño 3.4 index to define the ENSO event is pretty correct. Inevitably, there were some points stick together which will probably cause the classifier failure. In next stage, we want to know if we can push these points away more. Somehow, like we discuss in the section 2, the ISOMAP can reflect the real (geodesic) distance between the points.

In Fig. 3, the ISOMAP leading three components were shown. We found the ISOMAP points were indeed more separated than the points of the PCA. We noticed there were some events significantly different form others. They were the ENSO 82/83, 97/98, 15/16, and the La Niña 84/85, 88/89, 98/99. The 3D structure of ISOMAP had more variations in contrast to the PCA structure.

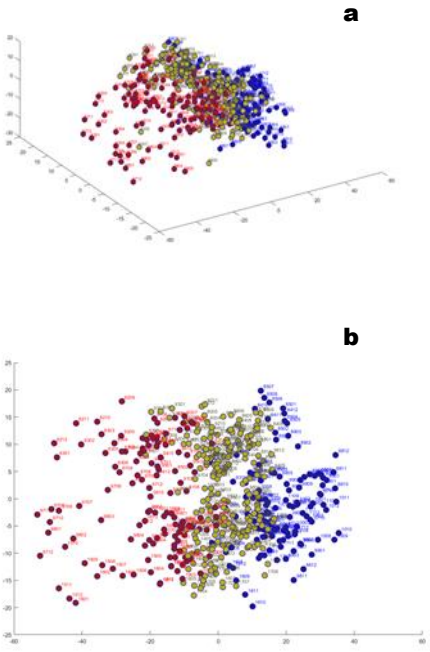

Fig. 2: (a) The 3D structure composed by the leading three PCA components. (b) the 2D structure composed by the leading two PCA components.

The explained variances of the PCA and the ISOMAP were similar. The first leading eigenvalue was occupied about 30% of the total sum of the eigenvalues. In here, we also check the residual variances, which is define as

$$1 - R^2(D_M, D_Y)$$

where R is Pearson correlation number, D_M can be the covariance matrix of the PCA or ISOMAP, and the D_Y can be the covariance matrix that comprises the low dimensional principal components.

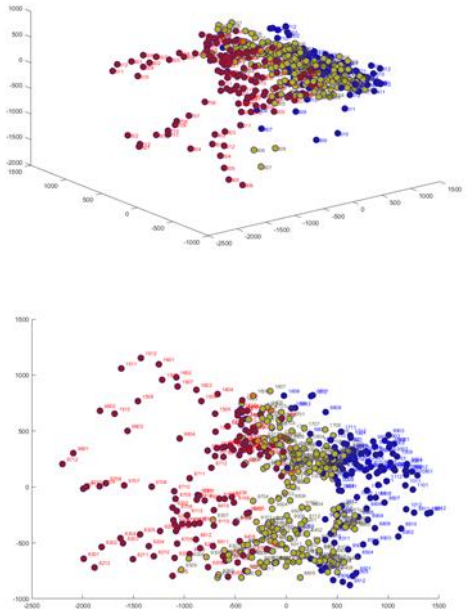

Fig. 3: (a) The 3D structure composed by the leading three ISOMAP components. (b) the 2D structure composed by the leading two ISOMAP components.

The residual variances reflected the similarity of the original covariance and the low dimensional covariance. The relation or the spatial/temporal structure between any two data points were kept in the ISOMAP calculation. The residual variances were shown in Fig. 4.

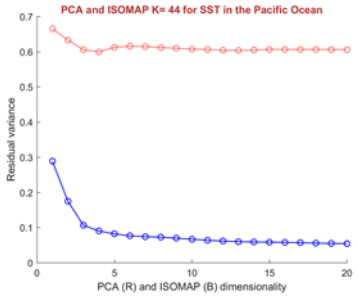

Fig. 4: The residual variances. The red curve is from the PCA and the blue curve is from ISOMAP.

When we checked the Fig. 4, we found the traditional PCA cannot reflect the original covariance well. In the leading 20 eigen components, there were at least 60% variance left. In contrast to the PCA, the leading three ISOMAP components almost extract 90% variance from the original covariance.

About the classification results from the leading 20 eigen components are similar in the PCA and the ISOMAP. The testing accuracy will be around 91% in both methods.

4. Conclusion and Discussion

The ISOMAP can help to identify the extreme ENSO cases and easily measure the differences between any two events. The residual variance of the ISOMAP is obvious lesser than the PCA. That implies we could rebuild the covariance matrix in low dimensionality. The ISOMAP results are easily to do the clustering. We all know there are no two identical ENSO events. But how different of them or what the quantity difference that it depends on the effective tool like the ISOMAP to measure. The clustering results can help us to check the ENSO cases. Besides the ENSO problem, the ISOMAP method also can use to do the composite analysis when we want to pick up similar cases in any Meteorological problems. The ISOMAP can be used to be diagnostic tool for checking different atmospheric circulations. We hope the ISOMAP can be one of the score measurements for checking the NWP outputs and the observation.

5. References

Alpaydin, E., 2014: *Introduction to Machine Learning*. MIT press, 640pp.

Bishop, C. M., 2006: *Pattern recognition and machine learning*. Springer-Verlag Press, 738pp.

Hsieh, W. W., 2009: Machine learning methods in the environmental sciences: Neural networks and kernels. Cambridge Univ. press, 349pp.

Ilin, A., H. Valpola, and E. Oja, 2006: Exploratory analysis of climate data using source separation methods. Neural Netowrk, **19**, 155-167.

Kelleher, J. D. and B. Tierney, 2018: *Data science*. MIT press, 264pp.

Lee, Y.-J. and O. L. Mangasarian, 2000: SSVM: a smooth support vector machine for classification. *Comput. Optim. Appl.*, **20(1)**, 5-22.

Russell, S. and P. Norvig, 2010: *Artificial intelligence: A modern approach*. Pearson press, 1132pp.

Tenebaum, J. B., V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, pages 2319-2323, 200.