Seasonal Forecasts using Data-Driven Models

John Chien-Han Tseng[2], Bo-An Tsai[2], Kaoshen Chung[2], Jiaching Wang[2]

[1] Central Weather Bureau
[2] National Central University

Abstract

The cost and the accuracy of numerical weather prediction (NWP) models for the long-term or so-called seasonal predictions are big issues. Computational fluid dynamics (CFD) numerical schemes are limited by appropriate time-integration steps for avoiding the instability in computing these NWP models. The time-integration steps are typically on the scale of seconds to hours, and the long-term prediction timescales are excess 30 days or several months. Other issues include the uncertainty of the proper initial condition, unresolved physical processes, the incomplete governing equation calculation, and numerical scheme errors increasingly over time. Traditionally, the operational weather centers will consider the ensemble members forecasts or modify the data assimilation cycle to response the unknown future. The simple way is to take the mean of NWP model predictions. Especially, when considering seasonal forecasts, for removing NWP model bias, the pentad, the 10-day, or monthly mean of predictions are usually used to obtain the final predictions. However, the cost of long-term NWP model integrations with average calculations is very high, and more than this, forecasts are still not good enough. One of the solutions to solve this dilemma is data-driven models, which do not have the limitation of using the short time-integration steps for the long-term seasonal forecasts. Moreover, data-driven models can combine different domain knowledge easily in learning algorithms for solving CFD problems. In term of computing efficiency, CFD problems can be solved in low-dimensional space instead of numerical schemes in an original high-dimensional space. The reasonable low-dimensional structures still can reflect the high-dimensional features, but the computation complexity in low-dimensional variables is simpler than the complexity in high-dimensional space. We proposed to combine the dimensionality reduction method, isometric feature mapping (ISOMAP), and numerical networks (NN) for Pacific sea surface temperature (SST) and East Asian precipitation predictions. The low-dimensional structures of SST and precipitation are shown by ISOMAP leading principal components (PCs). The evolution of leading PCs can be learned by NN algorithms, then the NNs predict these leading PC values. The predicted PCs with climatological residual PCs times climatological empirical orthogonal functions (EOFs) can be reconstructed SST or precipitation, which is the final prediction. Because the low-dimensional PC data points are less than the original physical data points, the NN calculation is faster than the traditional NWP models.

植基於資料驅動模式的季節預報

曾建翰[1]　蔡博安[2]　鍾高陞[2]　王家慶[2]

[1]中央氣象局　[2]國立中央大學

摘　　要

長期或所謂季節性預測的數值天氣預報（NWP）模型的成本和準確性是一個大問題。 計算流體動力學 (CFD) 數值方案受到適當的時間積分步驟的限制，以避免計算這些 NWP 模型時的不穩定性。 時間積分步驟通常為秒到小時，長期預測時間尺度超過 30 天或幾個月。 其他問題包括適當初始條件的不確定性、未解決的物理過程、不完整的控制方程計算以及隨著時間的推移逐漸增加的數值方案誤差。 傳統上，業務氣象中心會考慮集合成員的預測或修改數據同化週期以應對未知的未來。 簡單的方法是取 NWP 模型預測的平均值。 特別是，在考慮季節性預測時，為了消除 NWP 模型偏差，通常使用預測的五重、10 天或月平均值來獲得最終預測。 然而，長期 NWP 模型與平均計算相結合的成本非常高，除此之外，預測仍然不夠好。 解決這一困境的解決方案之一是數據驅動模型，它不存在使用短時間積分步驟進行長期季節性預測的限制。此外，數據驅動模型可以在解決 CFD 問題的學習算法中輕鬆結合不同的領域知識。 在計算效率方面，CFD 問題可以在低維空間中求解，而不是在原始高維空間中求解數值格式。 合理的低維結構仍然可以反映高維特徵，但低維變量的計算複雜度比高維空間的複雜度簡單。 我們建議結合降維方法、等距特徵映射（ISOMAP）和數值網絡（NN）來預測太平洋海面溫度（SST）和東亞降水。 海溫和降水的低維結構由 ISOMAP 主成分 (PC) 顯示。 領先 PC 的演變可以通過神經網絡算法來學習，然後神經網絡預測這些領先的 PC 值。 預測的 PC 與氣候殘差 PC 乘以氣候經驗正交函數（EOF）可以重建海表溫度或降水量，這就是最終的預測。 由於低維 PC 數據點少於原始物理數據點，因此 NN 計算比傳統 NWP 模型更快。

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.

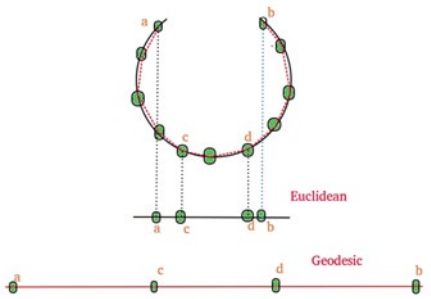

Fig. 1: The concept of the ISOMAP.

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.

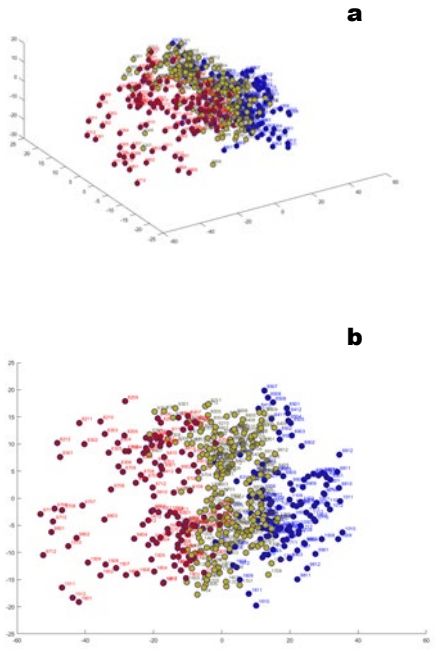

Fig. 2: (a) The 3D structure composed by the leading three PCA components. (b) the 2D structure composed by the leading two PCA components.

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.

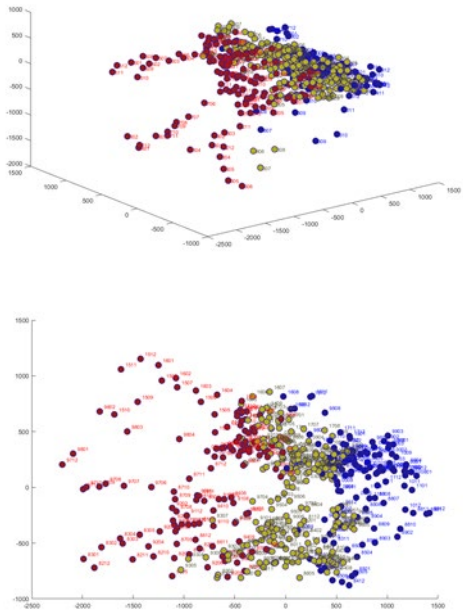

Fig. 3: (a) The 3D structure composed by the leading three ISOMAP components. (b) the 2D structure composed by the leading two ISOMAP components.

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.

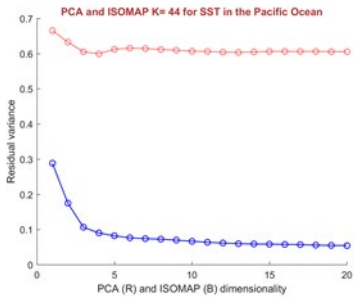

Fig. 4: The residual variances. The red curve is from the PCA and the blue curve is from ISOMAP.

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.