On the Advantages of Cluster Ensembles for the Detection of Characteristic Spatiotemporal Patterns

Mike Sips
Geoinformatics,
German Research Center
for GeoSciences GFZ
Telegrafenberg
Potsdam, Germany
sips@gfz-potsdam.de

Patrick Köthur
Geoinformatics,
German Research Center
for GeoSciences GFZ
Telegrafenberg
Potsdam, Germany
koethurp@gfz-potsdam.de

Abstract

Ensemble methods have been successfully applied to many real-world problems. Our intention is to raise awareness for a specific class of ensemble methods called cluster ensembles, and point to their advantages for the detection of characteristic spatiotemporal patterns. Our evaluation shows that the consolidated clustering result of the cluster ensemble is robust against outlying spatiotemporal patterns. Besides the analytical results, our collaboration with domain experts revealed two significant advantages of cluster ensembles for users. First, users do not make any strong assumptions on parameter assignments. Second, cluster ensembles lower the burden for users to utilize clustering algorithm in a proper way.

Keywords: Ensemble approach, Spatiotemporal Data, Detection of characteristic spatial and temporal profiles.

1 Introduction

Ensemble methods have been successfully applied to real-world problems in computer vision [7], computer security [2], medical diagnosis [8], credit card fraud detection [1] and to problems in many other scenarios. Our intention is to raise awareness for a specific class of ensemble methods called cluster ensembles, and point to their advantages for the detection of characteristic spatiotemporal patterns. The motivation behind this effort is that clustering is an important data analysis task for spatiotemporal data. It allows users to assess the data's spatial and temporal variability by focusing on the characteristic spatiotemporal patterns.

The detection of clusters in spatiotemporal data has an inherent difficulty. Clustering results often depend on specific parameter assignments users have to specify in advance. The derived clusters may favor certain aspects of spatiotemporal patterns. Hence, the detected clusters may not capture the characteristic spatiotemporal patterns.

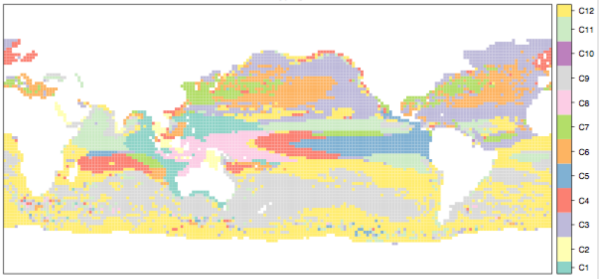
Cluster ensembles address this conceptual difficulty as follows. First, they compute many different clustering results for a given input data set. The different clusterings are computed by varying the parameter assignments of the utilized clustering algorithm; or they may even use different clustering algorithms. Second, they consolidate the different clustering results into a single result.

2 Experimental Setup

We explain the significant advantages of a consolidated clustering result in comparison to the result of a single clustering algorithm using sea surface data [6]. This data set contains thousands of spatially referenced time series; each time series describes the temporal behavior of a spatial position. We computed clusters of this data set according to the raw time series and to several statistical quantities, including the logarithmic power spectrum, standard deviation, minimum and maximum vectors of the time series. To detect clusters in sea surface data, we utilize the k-means clustering algorithm and vary the number of clusters for each statistical quantity between two and twelve. This produced 40 different clustering results. Note, we utilized the Euclidean distance in the computation of all clustering results.

These clustering results served as the input to our cluster ensemble. We computed the consolidated clustering result according to the rules described in [5]. Two ocean modelers assessed the 40 clustering results of the k-means algorithm as well as the consolidated clustering result.

Figure 1: Clustering result of the k-means algorithm on the raw data. The clusters are visually encoded using the color palette on the right hand side. This clustering result represents the *ENSO* processes well.

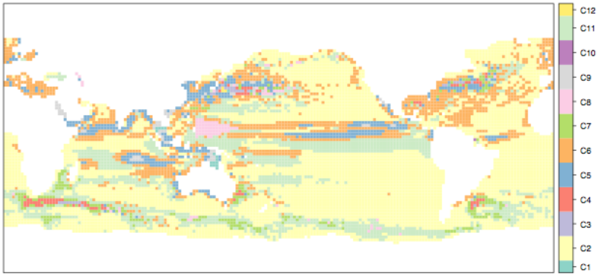


3 Results

This assessment showed that each of the 40 clustering results favored slightly different aspects of the sea surface data. Hence, these clustering results show a broad bandwidth of detected spatiotemporal patterns. After a detailed inspection of the detected patterns, it became apparent that many of these patterns do not represent real-world processes. In the following, we explain this finding in detail.

Figure 1 depicts the clustering result of the k-means clustering on the raw time series data. In this clustering result, the clusters *C4, C5* and *C8* represent different phases of the *El Nino and Southern Oscillation (ENSO)* processes, with high probability. Other processes such as the *Antarctic Circumpolar Current (ACC)* are not well represented in this clustering result. Figure 2 depicts the clustering result of the k-means clustering using the power spectrum. Here, the *ACC* is well represented since the clusters along the coastline of the Antarctic represent the *ACC* processes. However, the *ENSO* processes are not well represented in this clustering result.

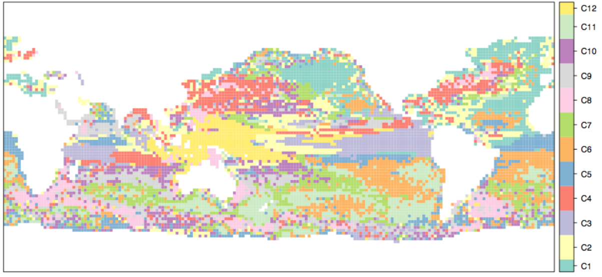
Figure 2: Clustering result of the k-means algorithm using the power spectrum of the time series. The clusters are visually encoded using the color palette on the right hand side. This clustering result represents the *ACC* processes well.



In contrast, the consolidated clustering result captures only the prominent spatiotemporal patterns of the input clusters. This result shows that the consolidated clustering is robust against outlying spatiotemporal patterns (see Figure 3). The consolidated clustering represents the *ENSO* and the *ACC* processes.

Our experiences in using cluster ensembles for the detection of patterns in spatiotemporal data are in line with the reported advantage of ensemble methods for non-spatiotemporal data [3] [4].

Figure 3: Consolidated clustering of the cluster ensemble. The clusters are visually encoded using the color palette on the right hand side. The consolidated clustering represents the *ENSO* and the *ACC* processes well.



Besides the analytical results, our collaboration with domain experts revealed two significant advantages of cluster ensembles for users. First, users do not make any strong assumptions on parameter assignments; for example, there is no need for users to guess the number of clusters in advance. Second, cluster ensembles lower the burden for users to utilize clustering algorithm in a proper way. The systematic variation of the statistical quantities and the parameter assignments strive to minimize the influence of misleading clustering results.

References

[1] P.K. Chan, W. Fan, A.L. Prodromidis, and S.J. Stolfo. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*, 14(6):67-74, 1999.

[2] I. Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone. Information fusion for computer security: State of the art and open issues. *Information Fusion*, 10(4):274-284, 2009.

[3] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993-1001, 1990.

[4] R. Shapire. The strength of weak learnability. *Machine Learning*, 5(2):197-227, 1990.

[5] A. Strehl and J. Gosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research,* 3:583-617, 2002.

[6] M. Thomas, J. Sündermann, and E. Maier-Reimer. Considerations of ocean tides in an ogcm and impacts on subseasonal to decadal polar motion excitation. *Geophysical Research Letters*, 28(12):2457-2460. 2001.

[7] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137-154, 2004.

[8] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S-F. Chen. Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine,* 24(1):25-36, 2002.