Clustering of Time Series using Wavelet Transformations as a Feature Extraction Mechanism

Kathy Norman, Ssurey Moon, Felix Huang, Josué Kuri
UCSC Extension. 30164:(003) Machine Learning and Data Mining

May 25, 2015

1 Introduction

A time series is a sequence of data points indexed by time at regular intervals. This model is used to represent a wide range of metrics such as the daily closing price of stocks, temperature, precipitation, population, etc. In the context of machine learning and data mining, clustering of a large set of time series is an exploratory technique aimed at identifying and understanding underlying patterns.

Considering every point of a time series as a dimension results in a high dimensional space which clustering algorithms cannot handle easily. These algorithms depend on a distance measure as a basis to maximize cohesion and separation. In ha high dimensional space, the contrast between the nearest and the farthest neighbor becomes smaller making it difficult for clustering algorithms to find meaninful groups [1].

Data dimensionality reduction is an approach to map a high dimensional space into a lower dimensional space such that the main characteristics of the data points in the original space are preserved and clustering on the lower dimensionality space results in meaningful groups. The two types of dimensionality reduction are feature selection and feature extraction. The former consists in selecting a subset of features from the original features. Feature extraction, on the other hand, generate a new set of features through some functional mapping.

Feature extraction techniques commonly used include Singular Value Decomposition (SVD), Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). Of these techniques, SDV is the most effective at reconstructing time series with minimal error. However, its time complexity $O(mn^2)$, where m is the number of time series and n is the lenght of each time series, makes this a computally-intensive approach [2]. A Fast Fourier Transform (FFT) algorithm can compute DFT coefficients in $O(mn \log n)$ and DWT, using a spacial type of wavelet called *Haar wavelet* can achieve $O(mn)$ [2].

In this project we use a feature extraction approach based on DWT using the Haar wavelet as the basis for the transformation. We create a generic framework for clustering of time series using this approach and apply the frame work to three types of time series: daily closing stock prices, daily values of exchange rates, and earthquake activity over time for various time regions.

2 Wavelet transformation

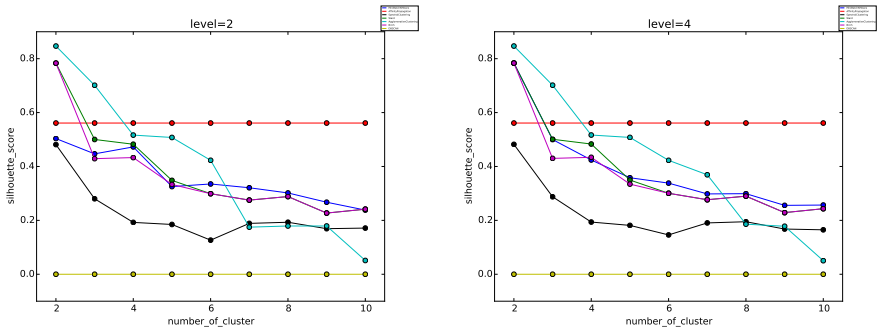
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3 Wavelet-based feature extraction

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

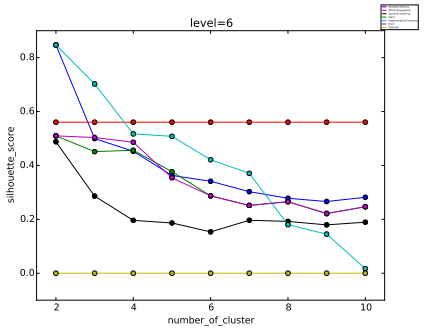
4 Experimental evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

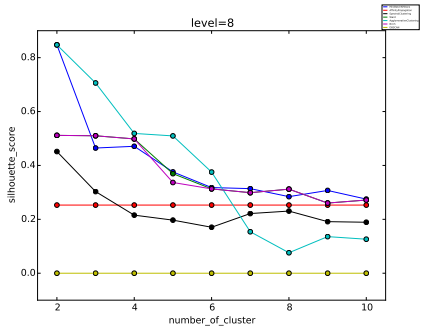


(a) Level 2

(b) Level 4

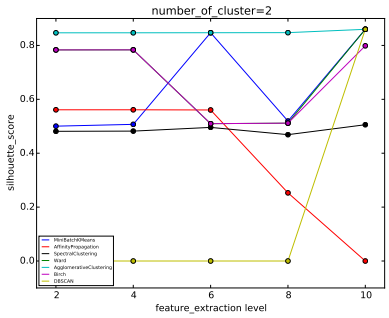


(c) Level 6

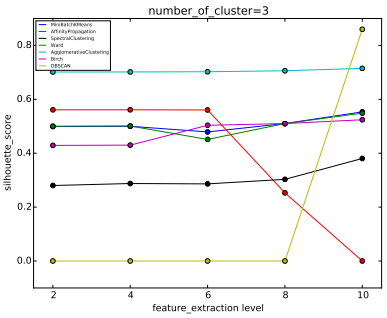


(d) Level 8

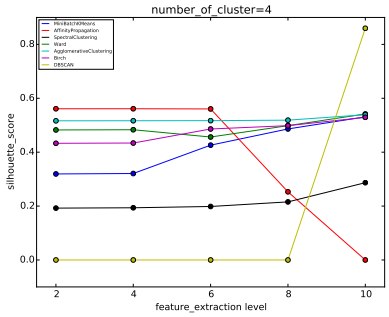
Figure 1: Silhouette score for various clustering levels.



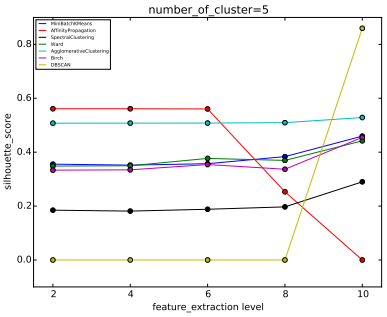
(a) Two clusters



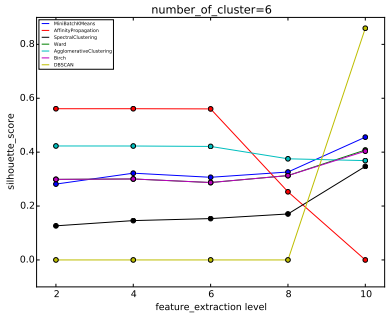
(b) Three clusters



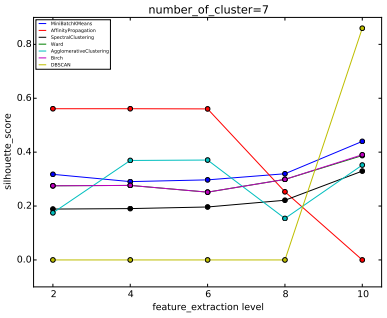
(c) Four clusters



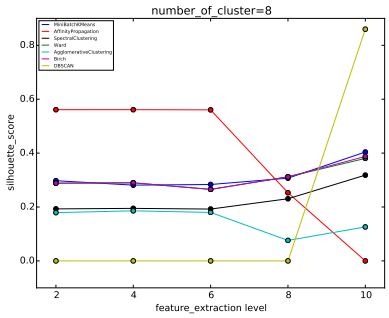
(d) Five clusters



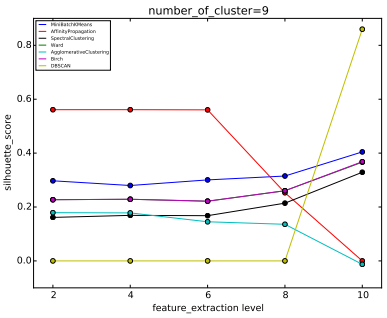
(e) Six clusters



(f) Seven clusters



(g) Eight clusters



(h) Nine clusters

Figure 2: Silhouette score for various clustering levels.

4.1 Evaluation criteria

4.2 Data description

4.2.1 Stock closing prices

4.2.2 Historic exchange rates

4.2.3 Historic earthquake data

4.3 Performance evaluation

5 Conclusions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

[1] K. Beyen, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, pp. 217-235, 1999.

[2] H. Zhang, T. B. Ho, Y. Zhang, M.-S. Lin. Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform In *Informatica*, Volume 30, pp. 305-319, 2006.