Fuzzy Clustering Time Series based on structural components

Ledys Llasmin Salazar Gomez*
Universidad de Valparaiso, Valparaiso, Chile - ledys.salazar@uv.cl

Rodrigo Salas
Universidad de Valparaiso, Valparaiso, Chile - rodrigo.salas@uv.cl

Abstarct: In the process of clustering time series main idea is to generate clusters with similar characteristics, so that greater accuracy in clustering is obtained. The present study describes the process of fuzzy clustering time series based on structural components. Techniques are implemented such as the Discrete Wavelet Transform (DWT) and Fuzzy c-means algorithm. The DWT allows mapping of the series in time domain and frequency and Fuzzy c-means algorithm allows the fuzzification of the series and obtaining clusters. The proposed algorithms are implemented on both synthetic data and real data from the actual data is considered pollution records of MP 10 for different times of day for 30 days. The clustering process, particularly the Fuzzy c-means, provides the ability to group time series according to the membership values thrown by the algorithm, where the maximum value determines cluster membership. Additionally, the process generates clusters characterized by a set of series belonging to the same stochastic process. Implementation, simulation and validation was performed on synthetic and real data, there was tested the hypothesis, finding that improves clustering algorithm on time series which have been transformed in the DWT in string comparison have not been transformed. Additionally, cluster validation indices realize the amount of time series must contain each cluster, so that the group is more accurate.

Keywords: Time series; Wavelets; Fuzzy c-means; cluster.

Introduction

In the framework of the Statistical there are different branches of knowledge that allow dealing with situations according to the conditions and assumptions about phenomena, such as: Sampling, Inference, Probability, Linear Models, Multivariate Models, Time Series, among others. Regarding the time series, the classic analysis is based on determining the components of the series and the autoregressive mode.

This paper is a contribution to the context of Time Series. A proposal for clustering time series based on fuzzy clustering algorithm is generated. The investigation is based on the implementation of an algorithm on synthetic and real data, which allows groups based on structural components and in turn supports validation of clusters through the different data rates.

In this regard, the time series grouping involves the formation of homogeneous groups, ie groups containing similar time series. Therefore, the importance of this process is the generation of clusters characterized as possessing a stochastic process with a high degree of accuracy.

Hypotheses and objectives

Hypotheses

This paper uses the fuzzy clustering algorithm Fuzzy c-means of sets of time series based on structural components. This algorithm allows to determine the cluster associated with the time series for the greater degree of membership that have regard to different clusters.

The clustering process improvement by making the clustering of time series based on structural components under the Discrete Wavelet Transform (DWT), compared to series that have not been transformed. Additionally, fuzzy clustering algorithm Fuzzy c-means generates time series group characterized by a given stochastic process.

Objectives

General objective

Develop a fuzzy clustering algorithm for time series based on structural components.

Specific objectives

- Adequate structural representation generate time series through the implementation of the DWT.
- Apply fuzzy clustering algorithm Fuzzy c-means on structural representations of time series.
- Generating fuzzy clusters on time series based on structural components.
- Perform the validation process through different cluster validation indices.

Theoretical foundations

Fuzzy Logic is based on the use of Fuzzy Sets associating a degree of membership of each of its elements, unlike conventional sets, where the elements belong or not to the whole. The notion of fuzzy set is mainly applied in the analysis of situations involving uncertainty related data, in turn, the respective modeling on them is structured according to the fuzzy rules used on the input data.

Recent studies suggest that the application of fuzzy rules on the inlet region are essential for modeling of the data. A related study interval length and which in turn implies the use of data mining techniques, was proposed by Chen, Hong Tseng in 2005 [1]. The approach uses a sliding window to generate continuous subsequences first of a series of time and then analyzes the sets of fuzzy elements of these subsequences. The final results are represented by linguistic rules and the method assumes that the membership functions are known beforehand.

The application of a set of fuzzy rules on the inlet region generates different time series models. On these models it is possible to make predictions that are determined by the components of the series. In this regard, Amjad, Jilani Yasmeen (2012) [2] propose an algorithm for predicting diffuse multivariate time series using the genetic algorithm and Particle Swarm. The techniques used in this algorithm are used for optimization, after this process the algorithm Taiwan Forex Exchange (TAIFEX) is applied to obtain better results and minimize the error rate compared to previous methods.

An article that illustrates the modeling of fuzzy time series was presented by Domanska and Wojtylak (2012) [3], they present a model to predict the concentrations of particles: PM10, PM2.5, SO2, NO, CO and O3. Using a distance function, actual data are compared with numbers to determine how the degree of membership. The pattern was prepared so that it can be used in the data usually imprecise and uncertain chaotic.

Therefore, it can be seen that the fuzzy modeling with all its elements is of wide applicability, as it provides the ability to model and interpret real-world situations involving data associated with uncertainty.

You can then say that the connection between logic and time series has become evident in recent decades, this joint has allowed the Data Mining Copper importance in different research in this area. Additionally, in this field modeling an adjunct to conventional modeling of time series, because it takes into account the stochastic process and the problem of the uncertainty in the data string is converted.

Another important reference that consolidated the research proposal has to do with Wavelet Transform.

In this regard Wong, Cheung, Zhongjie and Lui (2003) [4] studied the modeling and prediction of non-stationary time series through the use of Wavelets. The authors presented a procedure where the series as the sum of three separate components breaks down: trend, harmonics and irregular components. This method has been used for modeling the US dollar and is compared with other methods, where the results suggest that the prediction based on wavelet is a viable alternative to existing methods.

Fryzlewicz, Bellegem and Sachs (2002) [5] state that various studies involving longer time series variation in the background over time. They analyze the prognosis of non-stationary time series by Wavelets. Wavelets using stationary processes and introduce new predictors and in turn made a generalization of the Yule-Walker equations. Propose a method for automatic calculation of the choice of parameters and finally prediction algorithm, the prediction algorithm applied to a number of weather-related time.

Schluter and Deuschle (2010) [7] state that through the wavelet transform, a time series can be decomposed into a sum of time dependent frequency components. Through the wavelet transform for the seasonality is possible to capture their respective period and intensity over time, which implies that the prediction can be improved. In addition, they find that wavelets allow to improve the predictive quality of the data and eliminate noise that can occur in some autoregressive models.

Each of the studies mentioned above realize the importance of Wavelets processes in Time Series. The Wavelets allow recognition of the structural patterns of the series, so that its components can be studied.

In this paper it is assumed as a reference DWT, this transformation maps the signal in time domain and frequency and throws the scale and wavelet coefficients, for observing the characteristics of the series at different scales and frequencies.

Conclusions

This work is applied to synthetic and real data from the actual data, data associated with environmental pollution, especially for measurements of respirable particulate matter pollution MP10 are used. The initial time series measurements corresponded to the level of concentration in mg/m^3N of respirable particulate matter (PM10), between September 18 and October 17 of 2014, Pudahuel station monitoring.

Monitoring of environmental pollution levels plays a critical role in the health of human beings, and that higher levels of PM10 pollution, people are more likely to purchase common respiratory diseases such as bronchial asthma, obstructive bronchitis chronic pulmonary emphysema among others. Therefore, it is important to generate corrective and preventive actions to address this, because there are many factors that influence air quality, such as: Improper management of wastes, atmospheric conditions, solar radiation The vehicular pollution, among others.

In this work statistics and fuzzy techniques over different time series records containing environmental pollution per hour for 30 days were implemented. The implementation of these techniques yielded groups illustrating the similarity regarding PM10 pollution levels within 30 days. Therefore, it can be concluded that the algorithm Fuzzy c-means clustering allows for that are similar to each other on the level of contamination of MP10.

Additionally, it is worth noting that fuzzy clustering was performed on the data set containing the coefficients

of scale. The cluster on this dataset is based on consideration of the structural representation of the series under different levels of decomposition, which enables the mapping of the series in the time domain and frequency. The mapping of the series in this domain enables recognition of the cyclical and trend of the series, key components in the prediction process.

Importantly DWT implementation requires an adequate number of levels of decomposition. In this study, the simulation yielded four levels for the actual data.

The process of implementation of the DWT and the subsequent implementation of the algorithm of Fuzzy c-means, realize that it is possible to classify sets of time series not only from their autoregressive model, but also from transformed to realize its components.

Fuzzy clustering of time series based on structural components is more accurate than the clustering of time series that have not been transformed.
Finally, it can be concluded that the hypothesis was tested, donate the clustering process improvement to make a clustering of time series based on structural components under the DWT, compared to series that have not been transformed. Additionally, fuzzy algorithm Fuzzy c-means clustering groups generated time series characterized by a particular stochastic process.

References

[1] Chun-Hao, C., Tzung-Pei, H., and Vincent S. T., (2005). Analyzing Time-Series Data by Fuzzy.

[2] Usman, A., Tahseen A., J. and Farah, Y., (2012). A Two Phase Algorithm for Fuzzy Time Series Forecasting using Genetic Algorithm and Particle Swarm Optimization Techniques. International Journal of Computer Applications (0975 8887), 55 (16).

[3] D., Domanska and M., Wojtylak, (2012). Application of fuzzy time series models for forecasting pollution concentrations. Expert Systems with Applications.

[4] H. Wong, Wai-Cheung Ip, Zhongjie Xie and Xueli Lui, (2003). Modelling and forecasting by wavelets, and the application to exchange rates. Journal of Applied Statistics, Vol. 30, No. 5, 537553. The Hong Kong Polytechnic University, Hung Hom, Hong Kong and Department of Probability and Statistics, Peking University, Peoples Republic of China.

[5] P. Fryzlewicz, S. Van Bellegem and R. vonSachs, (2002). Forecasting non-stationary time series by wavelet process modelling. December 16. Universite catholique de Louvain, Institut de statistique, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium.

[6]

[7] Stephan Schluter y Carola Deuschle. Using Wavelets for Time Series Forecasting *Does it Pay Off*?. Department of Statistics and Econometrics, University of Erlangen-Nuremberg, Lange Gasse 20, Nuremberg.