Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 659–664

Procedia Computer Science

www.elsevier.com/locate/procedia

The 7th International Symposium on Emerging Inter-networks, Communication and Mobility (EICM)
August 9-12, 2020, Leuven, Belgium

Methodology for processing time series using machine learning

Noel Varela[a]*, Cesar Ospino[b], Omar Bonerge Pineda Lezama[c]

[a,b] *Universidad de la Costa, Barranquilla, Colombia*
[c] *Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

There are currently countless applications that can be cited in different areas of research and industry, where the data are represented in the form of time series. In the last few years, a dramatic explosion in the amount of time series ha occurred, so their analysis plays a very important role, since it permits to understand the phenomena described. A "time series" is a set of data of a certain phenomenon or equation, sequentially recorded. An alternative that allows to know the behavior and dynamics of a set of time series has been presented in the problem of classification, however, it is necessary to mention that most of the phenomena found in real life do not have a classification and that is why the unsupervised classification has brought great interest. Classification is organizing and categorizing objects into different, unlabeled classes or groups, which must be coherent or homogeneous [1][2]. This research proposes a methodology for obtaining the unsupervised classification of a set of time series using an unsupervised approach.

Keywords: Unsupervised classifier; Time series; Assembly of grouping algorithms.

1. Introduction

A time series is a set of numerical data, obtained from an experimental observation or by numerical calculation of

* Corresponding author. Tel.:+57-3235810446.
E-mail address: nvarela2@cuc.edu.co

10.1016/j.procs.2020.07.096

equations, i.e. a time series is a set of the form $ST = \{x1, x2, \ldots . xt, \ldots \ldots xN\}$ [3].

From the above, some outstanding characteristics should be considered in the analysis of the time series, which are described below [4][5][6]:

- Dimensionality: these are the degrees of freedom in the time series, that is, whether one is working in binary space, in real space, or in another.
- Size: is the amount of data that makes up a time series.
- Representation: a time series can be represented in a Cartesian plane, where the 'y' axis represents the value (or magnitude) and the 'x' axis is a consecutive index that corresponds to each value (which can be time or another variable) in the time series, therefore, the time series are in 1½ dimension.
- Structure: the time series contains peaks, which are not derivable or integrable.

One of the challenges in the classification of time series, is given by its structure, generally when classifying phenomena described by Attributes [7] [8] [9], their order does not affect the result, however, the time series conserves an order or temporality which does not allow to change the position of the data, reason why algorithms that work with attributes cannot be applied in this type of problems. Therefore, this paper presents a method of unsupervised and free form classification for time series, based on the development of different algorithms that are assembled to obtain the final grouping, with the purpose of revealing unknown objects/categories that help to a better understanding of the data, highlighting the inherent structure when grouping a set of time series.

2. Time series comparison

The comparison of time series is focused on the search for similarity of similar patterns, so in order to perform similarity analysis between two time series, it is necessary to quantify them, that is, assign them a numerical value. According to [10], "measurement is the act or process of assigning a number to a phenomenon, based on some rule".

The following definitions are given in [11]:

- "Similarity is an amount that reflects the strength or intensity of the relationship between two objects."
- "Distance is the measure of dissimilarity between two objects and refers to the discrepancy between two objects, based on various analyzed characteristics. It can also be interpreted as a measure of dissimilarity, between two objects".

In this study, distance techniques are used to give a similarity value between a pair of time series, so the relationship with similarity is given by: "Little distance equals Little difference, which equals Great similarity".

3. Assembly of grouping algorithms

By the No Free Lunch theorem [7], which tells that if an algorithm works well for a given problem, it will not have the same results for another problem, this paper presents different time series grouping techniques and applies an assembly algorithm to obtain the final grouping.

The clustering algorithm assembly [12] is generated by a set of clustering algorithms called "base clustering", which combines the outputs of the "base clustering" algorithms so that the useful information coded in each clustering algorithm is used to the fullest extent to obtain the final clustering.

Commonly, the assembly methods are applied mainly because they are able to drive weak algorithms and improve randomization [13].

In general terms, obtaining a cluster is relatively easy, since any partitioning algorithm generates a cluster, while the biggest difficulty lies in the combination of the algorithms, so to obtain success in the algorithm assembly, the key is in how the information given by the "base cluster" is expressed and how it is assembled [14].

4. Methodology

The proposed method for the grouping of time series considering an unsupervised and free approach is described as follows [15][16]:

1. A set of time series is selected.
2. A distance measurement is selected.
3. The proposed grouping techniques are applied.
4. The modified assembly algorithm is applied to the groupings obtained in step 3.
5. The final grouping is evaluated.

4.1 Data set

In this study, three data sets were selected, one of them containing synthetic data, one of them containing random data and one of them containing the transformation of 3D objects into 1D, where 1D corresponds to a time series [11].

4.1.1 Synthetic data set

To exemplify the most common problems of time series, a synthetic data set with 25 time series is used, which contemplates 3 different structures (three groups), which are Tables, Sine and Tables; to which different modifications were made to represent the problems of scale, lag, noise and combination of them.

4.1.2 Random data set

The normal distribution or Gaussian distribution is undoubtedly the most important and the other application in all continuous distributions, as it is quite adequate to describe the distribution of many data sets that occur in nature, industry and navigation, among others. For this reason, a data set made up of 180 time series, with normal distribution in a random way, was generated; with the objective of having a controlled data set, 9 groups were generated.

4.2 Data set of 3D objects

In Computing, the spatial data representation of a 3D figure is given by the definition of polygon mesh, which is very popular for three-dimensional models due to its simplicity. A data set was taken corresponding to the work done in [9], where they perform the transformation of 3D objects to 1D under the following idea "The 3D object is placed in a cube and given a predetermined order, the distance is recorded consecutively" (Figure 1).

The data set is made up of 5 classes that include: dolphins, dog, faces, cups and guns, with a total of 40 time series, where each time series has 1014 data.

4.3 Distance measurements

The distance measurements used and implemented in this study, contemplate the direct comparison using the Minkowski distance (equation 1) and indirect Fast Dynamic Time Warping or Evolved Fréchet [12].

Where x, y correspond to the series 1 and 2 respectively, k is the value contained in that position in the time series, n is the cardinality of the time series and λ is the order of distance to be calculated, if 1 corresponds to City Block distance, 2 is Euclidian and >= 3 Minkowski.

$$d = \sqrt[\lambda]{\sum_{k=1}^{n} |x_k - y_k|^{\lambda}}, \tag{1}$$

4.4 Grouping techniques

The proposed clustering algorithms are based on the search for representative time series given a data set, through the relationship between the representative time series and a measure of distance to all others. To form each of the groups found in the data set, a statistical cut-off criterion is used [11].

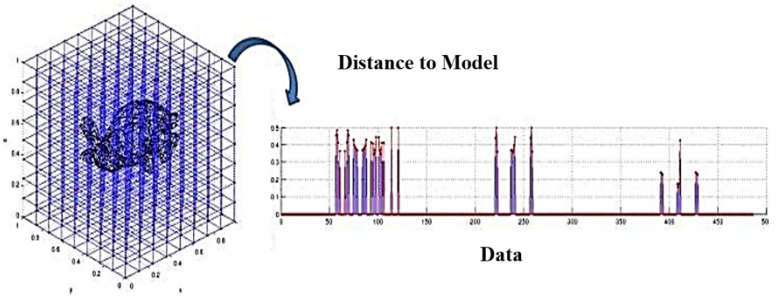

Fig. 1. 3D object data set, extracted from [9].

4.5 Grouping by representative series

For the grouping of time series, his paper focuses on the search for representative series, for which the following four criteria are considered [4]:

- Random: a time series is chosen at random.
- Minimum distance: the distance of each time series against all the others and the one with the minimum distance to some other.
- Maximum distance: the distance of each time series against all the others and the one with the maximum distance to some other.
- Centroid: for each time series, the distance to all the others is obtained and averaged, the series with the lowest average is taken as the representative series.

For each proposed criterion, three methods are presented below for time series grouping, using an unsupervised and free-form approach [1][3][14][16]:

4.5.1 Grouping method 1

This method considers the grouping of a given data set, using only a time series called a representative series.

1. Obtaining representative series: the representative series is taken depending on the desired criterion (random, minimum distance, maximum distance or centroid), the data set and its distance from all others.
2. Cut-off criterion: to generate the clusters, the distances of the representative series are ordered from the least to the most and the cut-off criterion is applied using the threshold set in 2.

4.5.2 Grouping method 2

This method modifies the cut-off criterion, since one could have groupings at the ends (one group or as many groups as time series in the data set), depending on the distribution of the data. Given method 1, step 2 is modified and step 3 is added.

1. Cut-off criterion: to generate the clusters, the distances of the representative series are ordered from less to more against the others and the cut-off criterion is applied using the threshold 0.5 (in each iteration the threshold will be increased by 0.5, until reaching 3.5).

2. Assessment: given the grouping with a cut-off threshold, it is assessed by means of Index I (the final grouping will be the one with the highest Index I).

4.5.3 Grouping method 3

This method examines the possibility of more than one representative series. Given method 2, step 2 and 3 are modified, and step 4 is added.

1. Cut-off criterion: to generate the clusters, the distances of the representative series against the others are ordered from lowest to highest and the cut-off criterion is applied using the threshold in 2.
2. Final grouping: from step 2, the first group formed is obtained and added to the final grouping, then these series are removed from the data set.
3. Subsequently, steps 1, 2 and 3 are repeated, until no time series exists in the data set.

4.5.4 Assembly method

Since in this work several algorithms are proposed for the grouping of time series, the disadvantage of having several groupings is presented, so it is necessary to implement a method to allow obtaining a single grouping. In this study, the modification of the assembly method by re-labelling is proposed [12], the modifications made are:

- The groups are not re-labelled.
- The cardinality of final groups is not fixed as the method, based on voting, does.
- Groups with cardinality >3 are examined.
- It is considered a group in the grouping, as long as it appears in the solution in more than 6 of the proposed techniques.

5. Experimental analysis

The 12 proposed grouping techniques were applied and then the assembly algorithm was applied to the three data sets, the results obtained are presented in Table 1. The data sets are also examined using the k-measurement algorithm, with the aim of having a reference for comparing results. In the results obtained, it can be seen that the proposed clustering method obtained better accuracy than the most k-medoids algorithm, regardless of the distance technique used. To exemplify one of the groupings obtained in Figure 2.

Table 1. Accuracy obtained using the algorithm assembly.

Distance techniques	Synthetic	Random	3D images
City Block	81.01	100	83.4
Euclidean	78.24	100	86
Minkowshi	81.36	100	57.4
FDTW	94.14	100	73.2
K-medoids	48.52	79	60.3

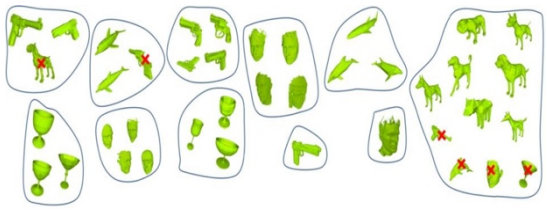

Fig. 2. Grouping of the 3D data set using the proposed assembly method and grouping techniques and the City Block distance measurement.

In Figure 2, the grouping of the data set of 3D objects is presented, where the 3D object that represents each time series is presented, as can be seen in the results, where the objects are grouped correctly, having only six objects badly grouped. Although it is not possible to obtain the 5 existing groups in the data set as the k-medoids algorithm would do, this does not affect the proposal, because the groups formed are consistent, that is, the groups contain elements that belong to the same class. It is worth mentioning that the grouping used was applying the transformation of these 3D objects to one dimension.

6. Conclusions

The problem of unsupervised classification of time series is to organize time series that are similar and to distinguish between those that are not. Considering the No Free Lunch theorem, this paper presents the grouping of time series under the unsupervised and free form approach, using twelve techniques that involve 4 different criteria and finally, to obtain a final grouping, the outputs of the proposed techniques are combined by modifying the assembly method by re-labelling. The results obtained by the assembly algorithm exceeds, in all tested cases, the k-medoids algorithm, which indicates the potential of the proposed method. In addition, one of the applications of this outstanding method is the grouping of 3D objects with 1D transformation, where besides having similarity in 1D in the groups formed when identifying the corresponding 3D object, these belong to the same category.

References

[1] Yan, S., Song, H., Li, N., Zou, L., & Ren, L. (2020). Improve Unsupervised Domain Adaptation with Mixup Training. arXiv preprint arXiv:2001.00677.

[2] Hunter, F. D., Mitchard, E. T., Tyrrell, P., & Russell, S. (2020). Inter-Seasonal Time Series Imagery Enhances Classification Accuracy of Grazing Resource and Land Degradation Maps in a Savanna Ecosystem. Remote Sensing, 12(1), 198.

[3] Yan, K., Huang, J., Shen, W., & Ji, Z. (2020). Unsupervised learning for fault detection and diagnosis of air handling units. Energy and Buildings, 210, 109689.

[4] Franceschi, J. Y., Dieuleveut, A., & Jaggi, M. (2019). Unsupervised scalable representation learning for multivariate time series. In Advances in Neural Information Processing Systems (pp. 4652-4663).

[5] Paris, C., Bruzzone, L., & Fernández-Prieto, D. (2019). A Novel Approach to the Unsupervised Update of Land-Cover Maps by Classification of Time Series of Multispectral Images. IEEE Transactions on Geoscience and Remote Sensing, 57(7), 4259-4277.

[6] Wang, S., Azzari, G., & Lobell, D. B. (2019). Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote sensing of environment, 222, 303-317.

[7] Viloria, A., Sierra, D. M., de la Hoz, L., Bohórquez, M. O., Bilbao, O. R., Pichón, A. R., … Hernández-Palma, H. (2020). NoSQL Database for Storing Historic Records in Monitoring Systems: Selection Process. In Advances in Intelligent Systems and Computing (Vol. 1039, pp. 336–344). Springer. https://doi.org/10.1007/978-3-030-30465-2_38

[8] Bode, G., Schreiber, T., Baranski, M., & Müller, D. (2019). A time series clustering approach for Building Automation and Control Systems. Applied energy, 238, 1337-1345.

[9] Ukil, A., Bandyopadhyay, S., & Pal, A. (2019, July). DyReg-FResNet: Unsupervised Feature Space Amplified Dynamic Regularized Residual Network for Time Series Classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8).

[10] Kim, H., Kim, H. K., Kim, M., Park, J., Cho, S., Im, K. B., & Ryu, C. R. (2019). Representation learning for unsupervised heterogeneous multivariate time series segmentation and its application. Computers & Industrial Engineering, 130, 272-281.

[11] Modak, S., Chattopadhyay, T., & Chattopadhyay, A. K. (2020). Unsupervised classification of eclipsing binary light curves through k-medoids clustering. Journal of Applied Statistics, 47(2), 376-392.

[12] Punmiya, R., Zyabkina, O., Choe, S., & Meyer, J. (2019, June). Anomaly Detection in Power Quality Measurements Using Proximity-Based Unsupervised Machine Learning Techniques. In 2019 Electric Power Quality and Supply Reliability Conference (PQ) & 2019 Symposium on Electrical Engineering and Mechatronics (SEEM) (pp. 1-6). IEEE.

[13] Ryabko, D. (2019). Time-series information and unsupervised learning of representations. IEEE Transactions on Information Theory.

[14] Yan, S., Song, H., Li, N., Zou, L., & Ren, L. (2020). Improve Unsupervised Domain Adaptation with Mixup Training. arXiv preprint arXiv:2001.00677.

[15] Viloria, A., Lis-Gutiérrez, J. P., Gaitán-Angulo, M., Stanescu, C. L. V., & Crissien, T. (2020). Machine Learning Applied to the H Index of Colombian Authors with Publications in Scopus. In Smart Innovation, Systems and Technologies (Vol. 167, pp. 388–397). Springer. https://doi.org/10.1007/978-981-15-1564-4_36.

[16] Pereira, J., & Silveira, M. (2019, February). Learning representations from healthcare time series data for unsupervised anomaly detection. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-7). IEEE.