PhD Thesis proposal:

A study on the discriminatory capacity of the temporal information on supervised time series classification problems

Amaia Abanda[1,2]
[1] *Basque Center for Applied Mathematics (BCAM)*
[2] *Intelligent Systems Group (ISG)*
Department of Computer Science and Artificial Intelligence,
University of the Basque Country UPV/EHU
Bilbao, Spain
aabanda@bcamath.org

Advisors: Usue Mori[2,3], Jose A. Lozano[1,2]
[3] *Department of Applied Mathematics*
Statistics and Operational Research
University of the Basque Country UPV/EHU
Bilbao, Spain

Starting date: 1 April 2017

***Abstract*—Time series classification has been always categorized as a particular case of classification in which the input objects are ordered sequences. As such, the research community has assumed that specific methods are required for dealing with this type of data, without really analysing this hypothesis. Specific methods are usually computationally expensive or demand some semantic treatment of the series that may turn the method cumbersome. In this thesis time series classification is addressed from a new point of view: the discriminatory power of the temporal information. In other words, given a dataset, we want to analyse the relevance of the temporal information (the order of the elements of a series, for instance) for classification and explore in which cases the specific methods are necessary and why in depth. Departing from distance based time series classification, the goal of this thesis is to explore which are the temporal characteristic of the time series data in order to measure how relevant they are for classification.**

***Index Terms*—time series, classification, temporal information, discriminatory, distance based**

I. Introduction

Time series data that is constantly being generated in a wide variety of fields such as bioinformatics, financial fields, engineering, etc [1]. Time series represent a particular type of data due the intrinsic temporal nature they have. A time series is an ordered sequence of observations of finite length, observations that are usually taken through time but they may also be ordered with respect to another aspect, such as space [2]. The large increase in the amount of this kind of data has given rise to an growing interest on studying and mining of time series.

In particular, this thesis is focused on time series classification (TSC), a data mining task which aims at finding a mapping between inputs (time series, in this case) and outputs (class labels o categories) based on some input-output pairs from where the model is learnt. In this manner, the goal of TSC is to predict the class labels of new unlabeled time series.

TSC differs from traditional classification problems in the ordered nature of the input objects. In traditional classification problems, the input objects are feature vector and the order of their attributes is not relevant; any re-ordering of them will produce the same results. In TSC, on the contrary, the input are time series, that is, ordered sequences. In this context, it is generally assumed that the order of the attributes is a discriminatory aspect and that, hence, time series can not be treated as feature vectors. As such, many specific methods for time series haven been proposed over the years.

The methods proposed for TSC can be divided into three main categories [4]: feature based, model based, and distance based methods. In feature based TSC, some representative features of the series are extracted and the series are transformed into feature vectors. Then, any traditional classifier can be employed, since the feature vector is not an ordered input. Some of the most typical feature based methods are discrete Fourier transform (DFT) [5] or discrete wavelet transform (DWT) [6]. In model based TSC, the main assumption is that the series belonging to the same class are generated by the same underlying model, so the prediction of a new series is done finding the model that best fits it. Some examples include auto-regressive models [7][8] or hidden Markov models [9]. Lastly, in distance based time series classification, the classification is done based on the concept of similarity. In this manner, the similarity between two series is defined employing a distance measure; those series that are close to each other respect to a given distance are considered similar, and those which are far away are considered different. It is assume that series which are similar in some sense, belong to the same class and this similarity concept is measured by using some specific distance measures. Then, this relationship is employed for classifying the new series. This thesis departs from distance based TSC.

Until recently, most of the work done in distance based TSC has focused on defining new distance measures and exploiting them within k-Nearest Neighbour classifiers (in particular, the 1-NN classifier). Different types of similarities focus on different aspects of the time series, such as the the shape, the temporal alignment, the correlation or the structure, so the simple approach of 1-NN has achieved a reputation of being difficult to beat [3][11]. However, this competitiveness may come from the strengths of the distance measures more than from the classifier itself [12]. As such, in the past few years many methods have been proposed that include the existing time series distances within more complex classifiers [13] [14] [15].

Time series distance measures can be categorized using many criteria. For instance, a widely employed categorization divides them into lock-step measures and elastic measures. Lock-step measures refer to those distances that compare the ith point of one series to the ith point of another (Euclidean distance, for example). Elastic measures, on the contrary, create a non-linear mapping between the series with the purpose of aligning them (Dynamic Time Warping [10], for instance). As such, elastic measures allow a comparison of one point of a series with another point of another series which is not at the same time instant. In other words, lock-step measures do not consider the relationship between an ordered subsequence of a series and another series, while elastic measures do. In this way, if a lock-step distance is employed in a classification task, the series are treated as feature vectors in which the order of the elements is meaningless. By contrast, if a elastic distance is used for classification, the main assumption is that the order of the elements of the series is a relevant aspect for discriminating between classes.

This dichotomy -whether the order of elements of the series is discriminatory or not-, has arisen from distance based TSC but it is, indeed, a more general matter. Most of the methods proposed until now assume that the temporal information of the series is relevant; the methods extract temporal features of the series in order to capture their temporal characteristics, but without really analysing the discriminatory power of these characteristics. As such, most of the existing approaches for TSC have assumed that, due to their temporal nature, specific methods are required for classifying time series. For instance, in distance based time series classification, this specificity is represented by the methods employing elastic measures.

The main problem with specific methods is that they are usually computationally expensive. In the case of distance based methods, for instance, most of the elastic measures take $O(n^2)$, where n is the length of the longest time series in the database. For large datasets, methods employing this kind of measures become cumbersome and unrealistic for a real application. Generic methods, on the contrary, can be computationally reasonable even for large datasets. In addition, specific methods have often a semantic part that the researches need to handle; for instance, in feature based method, the researches should know in advance which temporal features are more discriminatory for the given dataset. As such, the main motivation of this thesis is to understand beforehand in which cases the specific methods are necessary and in which cases a non-specific method would be enough. Hence, the researches could benefit from this knowledge to save time and effort on TSC. In addition, this novel point of view opens a new direction of research that has never been explored.

In the next sections the objectives of the thesis, the proposed methodology to carry them out, and the relevance of the present thesis are described.

II. OBJECTIVES

The main objective of this thesis to establish whether the temporal information of the time series -understood as the main characteristic of this type of data that differentiates it from other type of data- is, or is not, a discriminatory aspect for classification. Departing from distance based TSC, the goal of this thesis is to understand when specific methods are required or not (preferably in advance), and especially, why. This goal can be broken-down into the following sub-objectives:

1) **A taxonomy of distance based TSC**. As previously mentioned, many new distance based methods have arisen in the last years, in addition to the classical 1-NN. As such, this sub-objective aims at presenting a taxonomy that integrates all distance based approaches, structuring them in such a manner that the understanding of the field is improved. This sub-objective is in process at the moment and a review on distance based TSC has been recently submitted [2], which includes the proposed taxonomy.

2) **Analysis of need of specific methods within distances based TSC to develop a method that decides in advance whether specific methods are needed or not.** The categorization of the methods in specific/non-specific comes easily within distance based TSC by the differentiation of the methods which employ lock-step measures and elastic measures. Lock-step measures treat the series as feature vectors and, thus, are not specific for time series, while elastic measures take

into account the order of the elements of the series and, hence, are specific for time series. In this sub-objective, the temporal information is understood, as a first approximation, as the relevance of the order of the elements. Departing from the lock-step/elastic measures categorization - which no-consider/consider the temporal information, respectively- the main goal is to check if in those cases in which the temporal information is not discriminatory, non-specific methods perform better than specific. In this way, the idea is to develop a methods that decides in advance, for a given dataset, where elastic distance based or lock-step distance based methods will perform better. This sub-objective is in an early stage but some work has been already done in this direction [16].

3) **Analysis of need of specific methods in general TSC to develop a method that decides in advance whether specific methods are needed or not.** Following the idea of the previous sub-objective, the goal is to extrapolate the analysis of the specificity of the methods to general TSC. In particular, instead of considering the order as the unique temporal factor, new temporal characteristics will be considered. For instance, in feature based approaches, rather than extracting temporal features (as the amplitude of the series or their Fourier coefficients) , we could employ standard feature extraction techniques and compare both approaches in terms of accuracy. At the same time, we could check if there is a correlation between the performance of the specific/non-specific methods in distance based and feature based approaches, to somehow check if the weight of the order in the temporal information. The idea is to understand the intuition about which are the characteristic in a dataset that make the specific (or non-specific) methods more accurate.

4) **Proposal of a method that extract the temporal information's discriminatory power for TSC to know in advance which of the existing methods will be more appropriate for a given dataset.** The fourth sub-objective is to give a specific definition of the concept of *temporal information*. This definition will probably be broken-down into different characteristic as the order of the elements of the series or shift between series. The idea is to quantitatively measure, in advance, the power of these characteristics for classification. As such, one could extract these values, the discriminatory power, of each of the defined temporal characteristic for classification from a given dataset and employ them as a kind of weight. Then, the researcher can choose which of the existing method use in coherence with these weights. For instance, if the shift between series turn out to be very discriminative (large weight), a method that takes this fact into account should be used. In particular, most of these temporal weights could be small and that would mean that the temporal information is not relevant for classification. In this case, the non-specific methods would be probably more appropriate than the specific ones. To summarize, the goal is to develop a method that extract the weight of each temporal characteristic of a given dataset (in the sense how discriminatory they are for classification), in order to choose in advance which kind of method would be more appropriate.

5) **Application of the acquired knowledge to solve a real problem.** The goal is to give solutions to real time series classification problems by means of the development of a new method. The specific problem is: given a set of streaming time series with different kind of peaks, classify these peaks into different categories. Hence, it is a sub-series classification in a streaming context. An example of this scenario comes from water distribution companies, in particular, a company called Gipuzkoako Urak S.L. has proposed the following problem: this company has several sensors over a geographic region (Gipuzkoa) which measure the water flows passing through these localizations. These observations are taken over the time (every 5 minutes, for instance) and can be seen as streaming time series. In some specific time instants, there are high water consumptions (time series peaks), which may be caused by several reasons: someone filling the swimming-pool, fire-fighters taking water from a distribution point or water leakages. Since the companies do not differentiate between these peaks, a technician is send to the place every time that the water flow exceeds a established threshold. As such, most of times it is not a water leakage and the technician is sent for nothing. The objective is to propose a method that is able to classify these high water consumptions in the moment that they are happening in order to decide whether it is a water leakage or not (and thus, a technician has to be sent or not). In a general context, this problem can be seen as peak-including-sub-series classification in a streaming scenario.

   In the case of the problem proposed by the mentioned company, there is an additional challenge: they do not have the specific label of each peak, i.e., the reason of each high water consumption. Instead, they have some text comments written by the technician that went to check the specific high water consumption. In order to translate the problem to a TSC problem, we should have the specific labels. Hence, another goal within this objective is to apply some labelling technique to the comments in order to obtain a clear ground true of the existing categories.

III. Methodology and Work Plan

In this section, a general overview of the proposed methodology is presented. The work plan is defined in order to fulfill each one of the objectives of this thesis. For each objective, the goal is to develop a method (except for the first one: the taxonomy), which will be written and sent to a referred journal of the JCR. The preliminary findings may also be presented in conferences, preferably international. More concretely, the methodology for each objective is defined as follows:

1) **A taxonomy on distance based TSC:**
    a) A comprehensive revision of the existing distance based TSC methods, with emphasis on those that propose new ways of employing the existing time series distance measures.
    b) Extraction of the characteristics of each method in order to identify the possible categories of methods and explore the different criteria respect to which group the methods.
    c) Proposal of a taxonomy that structures the existing methods in a way that the possible future reader will have a better and more organized understanding of the distance based TSC methods.
    d) Writing of a review on distances based TSC based on the proposed taxonomy and submission to a JRC journal.
2) **Analysis of need of specific methods within distances based TSC to develop a method that decides in advance whether specific methods are needed or not:**
    a) State-of-the-art revision of any method that questions the use of specific methods for TSC, in particular, methods based on distances.
    b) Formulation of the hypothesis regarding the use of specific methods. In particular, a hypothesis that relates the distance based method employing lock-step/elastic (non-specific/specific) measures and general TSC methods that considers/non-considers the temporal nature of the series.
    c) Design of the experimentation and definition of the expected results.
    d) Contrast of the obtained results with the hypothesis to draw conclusions.
    e) Writing of a paper with the obtained conclusions. and submission to an international conference.
3) **Analysis of need of specific methods in general TSC to develop a method that decides in advance whether specific methods are needed or not.:**
    a) Revision of TSC methods in order to check if there are any methods that do not employ specific methods. Analysis of their justifications.
    b) Exploration of the different characteristic of the series (and in general, of the given dataset) that may determine when the specific/non-specific methods would obtain better/worse results and why.
    c) Formulation of the hypothesis about better/worse results in specific/non-specific methods depending on the different characteristic of the datasets.
    d) Definition of the experimentation and specification of the expected results and their correspondence with the formulated hypothesis.
    e) Interpretation of the obtained results and conclusions.
    f) Writing of a paper to present the obtained conclusions and submission to a JRC journal.
4) **Proposal of a method that extract the temporal information's discriminatory power for TSC to know in advance which of the existing methods will be more appropriate for a given dataset:**
    a) Revision of general classification methods which explore the discriminatory power of features obtained from the data. In particular, check if there is any similar work done for time series.
    b) Definition of temporal information and it's possible break-down into different characteristics. Definition of a way to measure the discriminative capacity of these characteristics.
    c) Proposal of a specific framework which, given a dataset, extract the temporal characteristics and the discriminatory power of each of them. Then, depending on the discriminatory power of them, gives a weight to each characteristics and choose the most appropriate method for classifying this dataset. In particular, if the temporal characteristics are found not to be relevant for classification, the proposed framework should choose a non-specific method.
    d) Definition of the experimentation to evaluate the proposed framework.
    e) Writing o a paper to present the developed framework and the analysis that underlies the proposal. Submission to a JRC journal.
5) **Application of the acquired knowledge to solve a real problem.**
    a) Definition of the problem with the help of the representative and the technician of the company Gipuzkoako Urak S.L.. The objective is to understand the structure of the data they have in depth and to define, with their help, the exact problem.
    b) Proposal of a labelling procedure to obtain the class labels of the high water consumptions from the comments of the technicians , as well as an evaluation method for the labelling procedure.
    c) Offline problem: split the time series in the time spans in which there is peack and classify these sub-series with the classes obtained in labelling procedure.
    d) Once the offline scenario is solved, provide a proposal to extrapolate the method to a streaming case.
    e) Presentation of the proposed method to the company to get a feedback.
    f) Writing of a paper with the proposed method and submission to a JRC journal.

IV. Relevance

Although the number of ways of employing a time series distance for TSC has increased over the years, a taxonomy that organizes the field has never been presented, so we think that it would be an important contribution for the researches in the community.

As previously mentioned, in almost every approach of TSC the proposed methods are specific for time series, assuming that the temporal information is a discriminatory aspect for classification. In our opinion, this hypothesis has to be analysed in detail, even if it is the base of practically all the existing approaches. The specific approaches for time series, specially those based on distances, require often high computational cost both for learning and prediction. In addition, depending on the length of the series and the size of the dataset, this methods can become unrealistically expensive in terms of computational time. In those that are not distance based, an specific treatment of the series is usually needed beforehand in order to extract the appropriate features or, in general, to fit an adequate model.

As such, we want to open a new direction of research that has never been explored: first, we want to specify the peculiarities of TSC respect to traditional classification problems from the point of view of the data analysis and, secondly, we want to measure these particular characteristics in terms of their discriminative power. In this manner, these characteristics could me measures in advance to know in order to choose the appropriate method to classify the given dataset. More specifically, these discriminative measures would help to understand whether it is necessary or not the use of specific methods. This progress would help the research community in a great manner by clarifying the relevant features of the time series data for the task of classification.

In addition, the acquired knowledge within this thesis will be applied to a real problem of time series peak classification in a streaming context. It is a innovative and challenging problem that combines different sub-problems of TSC but that has never been addressed specifically. An example where this method could be applied is in water distribution companies. Even if this challenge is a TSC problem, there are rather few works done by the researches in the field, so we think that proposing a comprehensive solution based on a novel methods will also be a innovation to the application domains of the field.

Acknowledgement

This research is supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through project TIN2017-82626-R funded by (AEI/FEDER, UE) and acronym GECECPAST. In addition, by the Research Groups 2013-2018 (IT-609-13) programs (Basque Government), TIN2016-78365-R (Spanish Ministry of Economy, Industry and Competitiveness). A. Abanda is also supported by the grant BES-2016-076890.

References

[1] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 102, 2002.

[2] Amaia Abanda, Usue Mori and Jose A. Lozano. A review on distance based time series classification. *Phttp://arxiv.org/abs/1806.04509*, 2018.

[3] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys*, 45(1):1–34, 2012.

[4] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40, 2010.

[5] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. *ACM SIGMOD International Conference on Management of Data*, pages 419–429, 1994.

[6] I. Popivanov and R. J. Miller. Similarity Search Over Time-Series Data Using Wavelets. *Proceedings 18th International Conference on Data Engineering (ICDE)*, pages 212–221, 2002.

[7] Anthony Bagnall and Gareth Janacek. A run length transformation for discriminating between auto regressive time series. *Journal of Classification*, 31(October):274–295, 2014.

[8] Marcella Corduas and Domenico Piccolo. Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, 52(4):1860–1872, 2008.

[9] Padhraic Smyth. Clustering sequences with hidden Markov models. *Advances in Neural Information Processing Systems*, 9:648–654, 1997.

[10] Donald Berndt and James Clifford. Using dynamic time warping to find patterns in time series. *Workshop on Knowledge Knowledge Discovery in Databases*, 398:359–370, 1994.

[11] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. *Proceedings of the 23rd ICML International Conference on Machine learning*, pages 1033—-1040, 2006.

[12] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, Reading, MA, addison-we edition, 2005.

[13] Rohit J. Kate. Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2):283–312, 2015.

[14] Arash Jalalian and Stephan K. Chalup. GDTW-P-SVMs: Variable-length time series analysis using support vector machines. *Neurocomputing*, 99:270–282, 2013.

[15] Pierre-François Marteau and Sylvie Gibet. On Recursive Edit Distance Kernels With Application to Time Series Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6):1–15, 2014.

[16] Amaia Abanda, Usue Mori and Jose A. Lozano. Requiere la clasificación de series temporales métodos específicos?. *Submitted*, 2018.