Multivariate Correlations for Early Classification

João Pedro Beirão
joao.beirao@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2018

Abstract

Multivariate time series are found in several data mining applications, where one or multiple variables are analysed over time. Early classification arises as an extension of the time series classification problem, in view of obtaining a reliable prediction as soon as possible. In temporal data, the order of the observations is critical, given that a particular sequence of occurrences can be the distinctive and discriminative factor. The same is true for early classification, where the goal is to use as less information as possible, ensuring a decent accuracy. The correlations between the variables of the multivariate time series from different time points may provide insights into predictive dependencies and relationships to be exploited. An information-theoretic method, named Multivariate Correlations for Early Classification (MCEC), for investigating the early classification opportunity from a collection of time series is proposed, implemented and assessed. Experimental validation tests are performed on synthetic, simulated and real data, confirming the ability of the MCEC algorithm to tradeoff between accuracy and earliness.
Keywords: early classification, multivariate correlations, Bayesian networks, model selection criteria

1. Introduction

Temporal data, generally known as Multivariate Time Series (MTS), consist of measurements or observations acquired and organized sequentially. In this context, one or multiple variables are examined over time, which means that the order of the information plays an important role. This sort of data is found in several data mining application areas such as medicine, economy, meteorology and marketing. Standard sequence classification involves using temporal data for constructing a classifier, which is able to predict the class label of a new given Time Series (TS), with a satisfactory accuracy. Early Classification (EC) can be viewed as an extension of the TS classification problem and it arises in scenarios where the anticipation of the prediction is beneficial. This matter has been a relevant subject of study in recent past, due to its several time-sensitive applications. The ability to obtain information in advance by having early knowledge about a specific event may be of great utility in many areas. For instance, a medical study [12] described how clinical data revealed that infants who were diagnosed with sepsis disease suffered from an unusual heartbeat twenty-four hours before the diagnosis. In this case, supervising the TS data of the infant's heartbeat and being able to classify it in advance, may lead to an effective early diagnosis and treatment.

In information theory, the idea of correlations between variables is associated to the analysis of the relationships and dependencies among them. In general, correlation can be viewed as a statistical technique for measuring how strong two random variables are related [18]. Moreover, a certain variable described over time through a TS is considered to be serially correlated if there is a statistical dependency between the values from different time periods [14]. Most real-world applications involve a degree of uncertainty, as a result of limitations in the information available and the challenges in modelling complex systems. Bayesian networks consist of probabilistic graphical models used for representing the information contained in a given dataset [18]. Their structures describe statistical dependencies and causal relationships between random variables. Data may contain unexpected correlations and their examination is useful for providing relevant knowledge to be explored, such as patterns and predictive associations. Considering the previously mentioned example of sepsis disease in infants, the investigation on the correlations among the clinical measurements and the patient's health condition was able to find a meaningful relationship.

This work aims to contextualize the EC problem, to explain its formulation and applications, and to review current algorithms on this matter. A method based on information theory is proposed, implemented[1] and assessed through experimental tests in synthetic, simulated and real data. This extended abstract includes only the experiments on the selected benchmark data.

2. Early Classification

Recent research has been focusing on EC. Seeing that earliness is intuitively related with temporal data, this

[1] https://github.com/joaopbeirao/MCEC-algorithm

problem deals with observations collected over time, generally referred to as TS. In this sort of data, the information is acquired and organized sequentially, which means that the order of the measurements has significance and their values are highly correlated. This is the case in electronic medical records, when the patient's health condition is monitored in each appointment and the information collected is structured chronologically.

Consider a dataset D as a collection of pairs:

$$(T_i, c_i) : i \in \{1, ..., w\}, \tag{1}$$

where T_i consists of a TS, c_i corresponds to its respective class label and w is the number of instances in D.

In general, a TS is defined as a vector of length L:

$$T_i = \left(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_L^{(i)}\right), \tag{2}$$

where each component: $\mathbf{x}_k^{(i)} = \left(x_{k_1}^{(i)}, x_{k_2}^{(i)}, ..., x_{k_N}^{(i)}\right)$, consists of N features measured at time point k.

In TS classification, a class label c_i is associated to each T_i through the relation: $Class(T_i) = c_i$. The standard TS classification goal is to construct a classifier from a training set, capable of assigning a class label to a new TS, with the maximum accuracy possible. Beyond optimizing the accuracy of the classification, in some applications it is beneficial to classify data as early as possible [29]. One of the fundamental challenges is the tradeoff between accuracy and earliness, since it is desirable to obtain a class label prediction without waiting for the end of the sequence, while ensuring an acceptable classification accuracy. EC of TS aims for making predictions as soon as enough data is available, and it is relevant in contexts where the collection of data has a cost associated or the delay of the predictions is adverse.

The work from Xing et al. [29] was one of the first to formulate the problem of EC. For a T_i, the subsequence:

$$t_i = \left(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}\right) \tag{3}$$

describes the section from the beginning until a time point $n \in \{1, ..., L\}$. This variable represents the timestamp from which the information of the TS can be neglected. An early classifier is able to find the time point n and perform an accurate classification based on t_i, such that: $Class(T_i) = Class(t_i) = c_i$. As stated by Xing et al. [30], it is important to distinguish EC from classic TS prediction, where the goal is to forecast values at some point in the future. EC of temporal data consists of anticipating the classification by using only a portion of the available information, without compromising the prediction quality. The goal is to predict the c_i of a T_i as early as possible given that the classification accuracy is close to the one using the complete data.

Literature complies with the fact that Rodriguez et al. [25] were the first to mention EC. They propose a TS classification method based on relative and region predicates that describe temporal intervals. Classifiers are constructed using these predicates, through a boosting approach, i.e. a method that generates ensembles of classifiers. This initial approach suggested the possibility of adapting boosting methods for addressing EC. Xing et al. [28] pointed that the standard sequence classification methods were only concerned with improving the accuracy of the classification. Hence, they studied the problem of EC on sequence data and proposed two methods: the Sequential Classification Rule (SCR) and the Generalized Sequential Decision Tree (GSDT). The goal consisted of finding sequence classifiers able to predict the class label of a new sequence, without using its entire length, while ensuring an expected accuracy. At first, in the SCR method, a set of features with effective characteristics for early prediction (frequency, distinctiveness and earliness) is extracted from the training data. Then, from the extracted features and based on both the expected accuracy and the prediction cost, a set of sequential classification rules is formed and used as the classifier. In the GSDT method, instead of an association rule, a decision tree is built, using a set of features as an attribute.

Moreover, Xing et al. [29] proposed an extension of an instance-based method based on the 1-nearest neighbour (1NN) classifier, with the Euclidean distance, for EC on TS data: the Early Classification on Time Series (ECTS) method. They identified two important requirements for an early classifier: being able to indicate the earliest time location of accurate classification; and ensuring an accuracy close to the case of using the full length TS. Thus, their approach involves a training phase, where the Minimum Prediction Length (MPL) is computed for each TS, yet based on a cluster of identical TS. This variable represents the time location (timestamp) from which the information of a TS can be discarded. Then, in the classification phase, for a new TS to be classified, if its 1-nearest neighbour from the training set has a MPL at most equal to the current timestamp being analysed, then the same class label is assigned to the new TS. In this work, the decision upon the tradeoff between earliness and accuracy is analysed, since the reliability of the 1NN is evaluated while anticipating the classification.

Considering some of the existing approaches for EC on TS and, in particular, the ECTS method, Xing et al. [30] identified one limitation: the interpretability. Based on TS shapelets, introduced by Ye et al. [31], the Early Distinctive Shapelet Classification (EDSC) method is proposed. This feature-based approach consists of extracting subsequences of TS (shapelets), which can distinctly point to the target class, and then selecting the ones more effective for EC. In spite of its effective performance, the proposed method is restricted to univariate time series (UTS) [10]. As an extension of EDSC, Ghalwash et al. [9] propose the Multivariate Shapelets Detection (MSD) method. The idea of EC on TS is maintained, however, it is generalized into a multivariate context. An N-dimensional shapelet is described as a set of multiple extracted subsequences, each of them asso-

ciated to one specific dimension. Likewise EDSC, a new TS is classified by finding the earliest covering shapelet from the generated subset. One drawback of shapelets pointed by Mueen et al. [23] is the significant computation time to extract them. Furthermore, He et al. [16] acknowledged the wide utility of EC on MTS and analysed the existing classification methods for this type of data. They identified two limitations of the MSD method [9]: the inability to extract time-independent subsequences from each dimension for the same multivariate shapelet, and the impossibility of dealing with dimensions of different length. As an attempt to address these issues, a new shapelet's quality evaluation approach is proposed by He et al. [16], as well as the Mining Core Feature for Early Classification (MCFEC) method. They introduce concepts related with the shapelets such as similarity degree, precision, recall and earliness; and they use them in the two steps of the MCFEC method: feature extraction and feature selection. Regarding the classification of a new MTS, He et al. [16] propose two methods for generating the classifier for early prediction of its class. The MCFEC-rule classifier, similarly to the SCR method [28], creates an association rule based on the selected core features (shapelets) from different dimensions. Concerning the other classification method, named MCFEC-QBC classifier, a Query By Committee (QBC) approach is used by matching the new MTS with the core features in order to find the predominant class. The methodology suggested by He et al. [16] focus on EC of MTS and it is flexible in dealing with the relevant information of distinct dimensions. In comparison with the MSD method [9], a significant progress is achieved in terms of the computation time of the training phase.

Parrish et al. [24] propose an approach based on a decision rule that uses linear or quadratic classifiers. The reliability threshold can be compared with the MPL parameter from the ECTS method [29], which is used to control the earliness of the classifier. The advantage of the parameter suggested by Parrish et al. [24] is the assurance on the reliability of the obtained decision, since the classification is performed only when the criterion is met. A similar measure is used in the model proposed by Ghalwash et al. [11], providing an uncertainty estimation of the predictions. Other solutions in the literature proposed multiple approaches for the EC problem. The work by Wang et al. [26] introduces the Earliness-Aware Deep Convolutional Networks (EA-ConvNets) method, which uses a neural network architecture to learn highly discriminative shapelets from TS data for making early class label predictions on incoming instances. In the work from He et al. [17], the issue of EC in imbalanced data is examined. Therein, the Early Prediction on Imbalanced Multivariate Time Series (EPIMTS) method, that uses an under-sampling technique, is presented. The work of Li et al. [19] proposes an approach for time-critical early decision making, that focus on modelling two aspects of MTS: temporal dynamics and sequential cues. Moreover, Hatami et al. [15] suggest a method based on a set of classifiers used sequentially in an iterative manner. Each classifier makes predictions with the portion of the TS available, but it also has a reject option in the case of an unsatisfactory classification. The methodology from Lin et al. [21] is called Reliable Early Classification (REACT) and it generalizes the EC study for MTS with numerical and categorical features.

One of the most recent approaches proposed for the EC on TS problem is presented by Mori et al. [22]. Similarly to the ECTS method [29], the accuracy and the earliness of the predictions are identified as the main objectives of EC on TS, and optimizing the tradeoff between both is perceived to be one of its fundamental challenges. As an attempt to tackle the problem of these two conflicting objectives, an EC method based on probabilistic classifiers is proposed: the ECDIRE [22]. They analysed some of the existing methods in the literature and developed an approach capable of dealing with three aspects simultaneously: avoid unnecessary calculations (specifically, forecasting and checking at all time points), control the reliability of the classifications (for instance, in the case of outliers), and measure the uncertainty of the predictions (a quantitative and interpretable evaluation).

3. Proposed Method

An introduction to some information and probability theory concepts is included in this section, in the interest of contextualizing the multivariate correlations methodology for the EC problem. Then, the proposed method is explained and the aspects concerning the implementation are described.

3.1. Information Theory

Information theory studies the transmission, processing, extraction and usage of information and it deals with a variety of information or communication sources, including the Discrete Memoryless Sources. These consist of independent random variables from a finite range of symbols (alphabet) and their respective probability distributions. *Entropy* quantifies the average uncertainty of a random variable. Considering the discrete random variable X, with a set of symbols (alphabet $\mathcal{X}$) and probability mass function $p(x) = P(X = x)$, where $x \in \mathcal{X}$:

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \tag{4}$$

Considering two discrete random variables X and Y and their joint probability $p(x, y) = P(X = x, Y = y)$, where $x \in \mathcal{X}, y \in \mathcal{Y}$, their *Joint Entropy* is defined by:

$$H(X, Y) = -\sum_{x} \sum_{y} p(x, y) \log_2 p(x, y). \tag{5}$$

Conditional Entropy measures the amount of information required to describe the outcome of X, given that the value of Y is known:

$$H(X|Y) = -\sum_{x} \sum_{y} p(x, y) \log_2 \frac{p(x, y)}{p(y)}. \tag{6}$$

Mutual Information quantifies the amount of information that one random variable gives about another:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \quad (7)$$

3.2. Probabilistic Graphical Models

Probabilistic graphical models attempt to describe the behaviour of complex systems using a graph-based framework for representing the probability distributions. *Bayesian networks* (BNs) are probabilistic graphical models for describing complex domains, and they can be used to represent the information about an uncertain system [18]. The BN representation consists of a directed acyclic graph G, characterized by a set of nodes $\mathcal{N} = \{X_1, X_2, ..., X_n\}$ and a set of directed edges E. Considering a $G = (\mathcal{N}, E)$, each node (vertex) corresponds to a random variable X_i, and the edges (arrows), that connect the nodes in a specific direction, describe the probabilistic dependencies between the random variables. For each node X_i, two sets can be defined: the set of parents Π_{X_i} and the set of non-descendants Φ_{X_i}. The structure of a BN is based on the assumption that each node X_i is conditionally independent of Φ_{X_i}, provided that Π_{X_i} is known [5]. The group of local probability models, representing the dependence of each variable X_i on Π_{X_i}, specifies the parameters for quantifying the network structure [4]. These form the set of conditional probability distributions $\Theta = \left\{\theta_{X_i|\Pi_{X_i}}\right\}_{i\in\{1,...,n\}}$, where:

$$\theta_{X_i|\Pi_{X_i}} = P(X_i = x_i|\Pi_{X_i} = \omega_i), \quad (8)$$

associated to each node X_i and conditioned on Π_{X_i}.

A BN $\mathcal{B} = (G, \Theta)$ is comprised of the direct acyclic graph structure G together with the set of parameters Θ. The joint probability distribution defined by this representation is calculated as [18]:

$$P_{\mathcal{B}}(X_1, ..., X_n) = \prod_{i=1}^{n} P_{\mathcal{B}}(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}}. \quad (9)$$

For a given dataset D, the problem of learning a BN consists of designing the $\mathcal{B} = (G, \Theta)$ that best represents D, according to a scoring function. The scoring function corresponds to the search guide for evaluating the effectiveness of the network in representing the data, and some of them are based on information theory concepts [5]. Moreover, when the structure of the network is fixed, the parameters Θ that maximize the scoring algorithms, for a given dataset, are those described by the observed frequency estimates [18]:

$$\hat{P}_{\mathcal{B}}(X_i = x_i|\Pi_{X_i} = \omega_i) = \frac{|D_{x_i,\omega_i}|}{|D_{\omega_i}|}, \quad (10)$$

for which $|D_{x_i,\omega_i}|$ represents the number of instances in D, where X_i takes the value x_i, and its parents (Π_{X_i}) take the value ω_i. Similarly, $|D_{\omega_i}|$ denotes the number of instances in D, where Π_{X_i} takes the value w_i.

The *Minimum Description Length (MDL)* principle is known as an Occam's razor approach to select, for a given dataset, the best fitting model and its parameters. It states that, for a certain data and a number of alternative models, the best option corresponds to the simplest model [7]. In the problem of learning a BN, the Bayesian Information Criterion (BIC) is known as the MDL score. It is concerned with analysing the tradeoff between the Log-Likelihood (LL) of the dataset D (the effectiveness of the fit to the data) and the complexity of the model. This scoring function is defined as [18]:

$$MDL(D|\mathcal{B}) = LL(D|\mathcal{B}) - \frac{\log_2 N}{2}|\mathcal{B}|, \quad (11)$$

where N corresponds to the size of the data, and $|\mathcal{B}|$ represents the model dimension (number of parameters in $\mathcal{B}$). The LL term quantifies the amount of information required to describe the dataset D, using $\mathcal{B}$. Conversely, the penalty term measures the amount of information needed to encode the model $\mathcal{B}$ [5]. It is desired the most effective fit to the dataset, provided that the complexity of the model is as low as possible.

Similarly to the MDL scoring function, the *Akaike Information Criterion (AIC)* [1] corresponds to a measure of the quality of statistical models for describing a given dataset. In the problem of learning a BN, the difference between MDL and AIC is associated to the penalty applied to the number of parameters $|\mathcal{B}|$. The AIC scoring function can be defined as [5]:

$$AIC(D|\mathcal{B}) = LL(D|\mathcal{B}) - |\mathcal{B}|. \quad (12)$$

In Equation (11), the second term quantifies the amount of information required to encode the model $\mathcal{B}$, where each parameter in the set Θ is considered to use $\frac{1}{2}\log_2 N$ bits. Conversely, in Equation (12) each parameter of Θ is considered to use 1 bit. This means that the penalization on the number of independent parameters is stronger in the MDL scoring function than in the AIC score. Likewise for the MDL score, the best model corresponds to the one that maximizes Eq. (12).

3.3. Multivariate Correlations

From a statistical point of view, the concept of correlation between variables attempts to measure the relationships and dependencies among them. The knowledge of how the variables are related, as well as of what inferences can be made about their causal relationships, is useful for drawing conclusions about potential predictive relationships to be analysed and exploited.

For a finite set of discrete random variables $S = \{X_i\}_{i=1,...,n}$, with joint probability distribution $P_S(X_1, ..., X_n)$, the total correlation between those variables can be defined as [7]:

$$I(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, ..., X_n). \quad (13)$$

In this case, the mutual information measures the dependencies among the variables, i.e. the amount of information that these quantities give about each other.

Let a structural relation R be a subset of the system S. Its joint probability distribution corresponds to the

marginal distribution from S:

$$P_R(X_{R_1}, ..., X_{R_k}) = \sum_{X_i \notin R} P_S(X_1, ..., X_n), \quad (14)$$

where k is the number of elements in R.

Definition 1. A structure over the system S with underlying joint probability P_S is a pair $(\mathscr{S}, P_{\mathscr{S}})$, where $\mathscr{S} = \{R_j\}_{j=1,...,k}$ is a collection of structural relations and $P_{\mathscr{S}}$ is a joint probability distribution over S s.t.:

1. No $R_i \in \mathscr{S}$ is contained in another ($\forall_{i,j}\ R_i \nsubseteq R_j$);
2. Every $X_i \in S$ is included in at least one $R_j \in \mathscr{S}$;
3. $P_{\mathscr{S}}$ is the solution to the optimization problem:

$$\max_{P \in \mathcal{P}} \quad H(P)$$

$$\text{s.t.} \quad \sum_{X_i \notin R_j} P_{\mathscr{S}}(X_1, ..., X_n) = \sum_{X_i \notin R_j} P_S(X_1, ..., X_n),$$

$\forall R_j \in \mathscr{S}$, where $\mathcal{P}$ is the set of probability distributions of the variables from S.

For example, from the set of discrete random variables $S = \{X_1, X_2, X_3, X_4\}$, some admissible structures $\mathscr{S}$ correspond to $\{\{X_1, X_2, X_3\}, \{X_4\}\}$, $\{\{X_1, X_2\}, \{X_3, X_4\}\}$ or $\{\{X_1, X_2\}, \{X_1, X_4\}, \{X_2, X_3, X_4\}\}$. Conversely, $\mathscr{S} = \{\{X_1, X_2, X_3\}, \{X_1, X_3\}, \{X_4\}\}$ is not an acceptable structure since the relation between X_1 and X_3 is included in two structural relations, which represents a transgression of the first property. Similarly, $\mathscr{S} = \{\{X_1, X_2\}, \{X_2, X_4\}\}$ does not consist of a proper structure because the variable $X_3 \in S$ is not part of any structural relation from $\mathscr{S}$, as required by the second statement.

For a given system $S = \{X_i\}_{i=1,...,n}$ and an associated set of structural relations $\mathscr{S} = \{R_j\}_{j=1,...,k}$, the mutual information $I(S)$ represents the maximum amount of information that the variables X_i from S provide about each other. On the other hand, $I(\mathscr{S})$ quantifies the information described by the correlations inside the structural relations R_j. The difference $I(S) - I(\mathscr{S})$ measures the knowledge of the dependencies and relationships between the variables of S that are not included in the relations that compose $\mathscr{S}$. From Eq. (13), this value can be described a difference of entropies:

$$\begin{aligned} I(S) - I(\mathscr{S}) &= \sum_{i=1}^{n} H(X_i) - H(S) - \sum_{i=1}^{n} H(X_i) - H(\mathscr{S}) \\ &= H(\mathscr{S}) - H(S). \end{aligned} \quad (15)$$

Eq. (15) is always non-negative, because $H(S)$ consists of the lowest possible average number of bits required to describe the random variables from S. Similarly, this difference represents the information given by the existing correlations in S, that is not incorporated in $\mathscr{S}$.

3.4. MCEC algorithm

Consider a TS T, as in Eq. (2), representing the evolution of the variable X over time, and its respective class label C, acting as another variable correlated with T. The set of X_k can be viewed as a collection of time dependent discrete random variables, for which a joint probability distribution can be defined. The correlation between any two variables measures the influence that the value of X at one time point has on the value of X at another instant. Note that, since a TS is chronologically organized, it is relevant to analyse the dependency of variables on their early states, i.e. the degree of dependence of X at a certain time point on the value observed at a previous instant. Similarly, the correlation between C and X_k quantifies the influence that the variable X at time point k has on the class label. In the EC context, the focus is to study systems where the class labels verify a high dependence on a certain amount of early states of X_k, while the remaining time points are dispensable for a satisfactory classification.

Consider the finite set of discrete random variables S to be composed of the TS T together with its respective class label C. The system, with an associated joint probability distribution $P_S(X_1, X_2, ..., X_L, C)$, where L represents the TS length, is defined as:

$$S = \{X_1, X_2, ..., X_n, X_{n+1}, ...X_L, C\}, \quad (16)$$

for which n corresponds to a specific instant in the TS, designated early time point. The goal is to find the value n that describes $P_S(X_1, X_2, ..., X_n, C)$ such that:

$$P_S(C|X_1, X_2, ..., X_L) \approx P_S^n(C|X_1, X_2, ..., X_n). \quad (17)$$

The conditional probability $P(X|Y)$ measures the likelihood of the event X, given that the event Y is observed. Therefore, $P_S^n(C|X_1, X_2, ..., X_n)$ and $P_S(C|X_1, X_2, ..., X_L)$ describe the probability of the class label C occurring, provided that some or all variables of T are known, respectively. Seeing that $n < L$, Eq. (17) denotes that the variables $\{X_1, X_2, ..., X_n\}$ characterize the class label of the TS almost as accurately as using the entire T. In addition, a criterion is required for identifying the optimal n for which the complexity of the model defined by P_S^n is low, provided that the majority of correlations from P_S are considered.

In general, sequence classification methods are performed in a collection of T_i with their respective C_i, organized in a dataset D (Eq. (1)). In some cases, the joint probability distribution P_S is not known in advance, thus it has to be computed from the data, through maximum likelihood estimation. In particular, given a dataset D, with size w, as the system S, the distribution P_S that maximizes the likelihood of D is such that:

$$\hat{P}_S(X_1 = x_1, \dots X_L = x_L, C = c) = \frac{|D_{x_1,...,x_L,c}|}{w} \quad (18)$$

for which $|D_{x_1,...,x_L,c}|$ is the number of instances in D, where each X_i takes the value x_i and C the value c.

Given the system S, described in Eq. (16), the set of structural relations, defined by:

$$\mathscr{S}_n = \{\{X_1, ..., X_n, X_{n+1}, ..., X_L\}, \{X_1, ..., X_n, C\}\}, \quad (19)$$

depends on the value of n and it corresponds to a structure that respects the previously described properties. Considering $A_n = \{X_1, ..., X_n\}$ and $B_n = \{X_{n+1}, ..., X_L\}$, the structure is represented as:

$$\mathscr{S}_n = \{\{A_n, B_n\}, \{A_n, C\}\}. \tag{20}$$

The structural relation A_n contains the information about the evolution of the variable X until the time point n, i.e. the early states of the collection of TS. On the other hand, B_n describes the remaining instants of T_i which can be viewed as the knowledge about the later states of the variable X. Finally, C represents the class label information from the collection of TS. The structure $\mathscr{S}_n$ can be seen as a simplified model of the system S. It is expected to include the correlations between the early and the later information about the TS (A_n and B_n), as well as between the early states of T_i and the knowledge about their classes (A_n and C). Conversely, the correlations between B_n and C are not preserved because the idea is to study the possibility of describing the class from the early states A_n, while neglecting the information from B_n. The probability distribution of $\mathscr{S}_n$ is obtained based on Theorem 1 and considering the BN represented in Fig. 1.

Theorem 1. Consider the BN $\mathcal{B}_n = (G_n, \Theta_n)$ with G_n given by Fig. 1 and Θ_n calculated according to Eq. 10. Let $\mathcal{B}_n$ represent D as the system S, with underlying probability given by $\hat{P}_S$, as in Eq. (18). The structure $(\mathscr{S}_n, P_{\mathscr{S}_n})$ over S has a probability distribution equal to the joint probability distribution of $\mathcal{B}_n$, i.e., $P_{\mathscr{S}_n} = P_{\mathcal{B}_n}$.

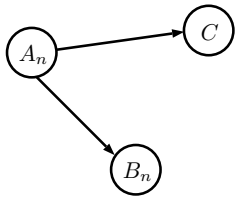

Figure 1: BN representation of $\mathscr{S}_n$ over the system S.

Thus, from $\Pi_{A_n} = \emptyset$, $\Pi_{B_n} = \{A_n\}$ and $\Pi_C = \{A_n\}$, and through Eq. (9):

$$P_{\mathscr{S}_n} = P(A_n)P(B_n|A_n)P(C|A). \tag{21}$$

From Eq. (15) and for each value of n, the difference of entropy applied to these context is represented as:

$$\begin{aligned} I(S) - I(\mathscr{S}_n) &= H(\mathscr{S}_n) - H(S) \\ &= H(C|A_n) - H(C|A_n \cup B_n). \end{aligned} \tag{22}$$

The conditional entropy is used to quantify the uncertainty about the classes of the collection of TS, given that T_i is fully or partially known. On the one hand, $H(C|A_n)$ consists of the amount of information required to predict the class labels, provided that the TS are known until the time point n. On the other hand, $H(C|A_n \cup B_n)$ corresponds to the amount of information needed to describe C_i, based on the knowledge of the entire T_i. The difference between these two conditional entropies measures the knowledge that the whole TS provides about the classes (correlation between C and $A_n \cup B_n$) which is not represented by the incomplete data (correlation between C and A_n). Thus, Eq. (22) can be viewed as the lack of information caused by describing the structural relation C from A_n, i.e. the loss of knowledge for using the collection of TS only until the early time point, in the classification process.

In addition to earliness in predicting the classes, the goal consists of finding the value n for which $\mathscr{S}_n$ represents the system S with a reasonable complexity. Since this can be seen as a problem of learning the BN from Fig. 1, both MDL and AIC scores are applied to the multivariate correlations for EC approach, in the interest of finding the best fitting model. These scores are used as two criteria for choosing the early time point, such that the selection of the model takes its simplicity into consideration. From Eq. (11) and considering $P_{\mathscr{S}_n}$, described in Eq. (21), the MDL score is defined as:

$$MDL(D|\mathscr{S}_n) = \frac{\log_2 w}{2}|\mathscr{S}_n| - \sum_{i=1}^{w} \log_2[p(C|A_n)p(B_n|A_n)p(A_n)], \tag{23}$$

where w is the number of instances in the dataset D, $|\mathscr{S}_n|$ denotes the number of independent parameters in the model, and $P_{\mathscr{S}_n}$ is associated to $\mathscr{S}_n$, which describes S as a representation of the given data. Similarly, the AIC score, applied to this context, is defined as:

$$AIC(D|\mathscr{S}_n) = |\mathscr{S}_n| - \sum_{i=1}^{w} \log_2[p(C|A_n)p(B_n|A)p(A_n)]. \tag{24}$$

As represented in the direct acyclic graph structure from Fig. 1, the goal is to analyse how the structural relation A_n is able to describe C, while the correlation between B_n and C is neglected. For this reason, the computation of the network complexity only considers the relation between the early states and the class labels:

$$\begin{aligned} |\mathscr{S}_n| &= |\{A_n, C\}| = ||A_n|| - 1 + (||C|| - 1)\,||A_n|| \\ &= ||A_n|| \times ||C|| - 1, \end{aligned} \tag{25}$$

where $||A_n||$ and $||C||$ denote the number of distinct observations in the structural relation A_n and C, respectively. In Eqs. (23) and (24), the first term quantifies the complexity of the model, i.e. the amount of information required to encode not only $\mathscr{S}_n$, but also the data given $\mathscr{S}_n$. The second term measures the LL of the data based on the model, i.e. the amount of information needed to represent the dataset D according to the probability distribution $P_{\mathscr{S}_n}$. While n increases, the size of A_n becomes larger, the number of correlations is higher and, consequently, the complexity of the model increases. In addition, the more information about the TS there is, the better the correlations describe the data, which means

a decrease in the number of bits needed to describe C from A_n. The difference between these two terms describes the tradeoff between the model complexity and the effectiveness of the fit to the data. Note that Eqs. (23) and (24), are the symmetric of the definitions described in Eqs. (11) and (12), respectively. This means that, the best model is the one that minimizes the scoring functions. The simplest model, that is able to use the least amount of correlations while it maintains a distribution as close to the original as possible, is found through minimizing both $MDL(D|\mathscr{S}_n)$ and $AIC(D|\mathscr{S}_n)$.

3.5. Implementation

The proposed algorithm is implemented in Java language, using some functionalities of Weka Data Mining Software [13]. The Multivariate Correlations for Early Classification (MCEC) program, summarized in Algorithm 1, receives as input a comma-separated values (CSV) file, containing the TS and the respective class labels. The number of attributes is also required as input. Both UTS and MTS are allowed, however, the TS must be of fixed length. In the interest of verifying the reliability of this EC approach, an investigation on the performance of multiple classifiers is done, while varying the length of the TS. Seven classifiers are considered (Table 1), using the default parameters and stratified cross-validation with 10 folds. The outcomes of the difference in entropy, LL, MDL score, AIC score and classification accuracy, all for $n \in \{1, ..., L\}$, are outputted in text files. An additional Matlab script is provided for generating the five graphs with the results.

Algorithm 1 MCEC program.

1: **for** $n \in \{1, ..., L\}$ **do**
2: Separation of data from D in five groups: $\{A_n\}$, $\{C\}$, $\{A_n, C\}$, $\{A_n, B_n\}$ and $\{A_n, B_n, C\}$
3: Count number of occurrences of each case in each group
4: Calculate the probability values: $P(A_n = a)$, $P(A_n = a, B_n = b)$, $P(A_n = a, C = c)$ and $P(A_n = a, B_n = b, C = c)$
5: Compute $H(C|A_n) - H(C|A_nB_n)$
6: Count number of independent parameters: $||A_n||$ and $||C||$
7: Compute $|\mathscr{S}_n|$ and $LL(D|\mathscr{S}_n)$
8: Compute $MDL(D|\mathscr{S}_n)$ and $AIC(D|\mathscr{S}_n)$
9: Compute classification accuracy with the TS until n
10: Output the five vectors

Classifier	Description
NB	Naïve Bayes
BN	Bayes Net
SMO	Sequential Minimal Optimization
J48	C4.5 decision tree
REPTree	Reduces Error Pruning Tree
RandFor	Forest of multiple random trees
kNN	k-Nearest-Neighbor

Table 1: Classifiers used for comparing with the results.

4. Results on benchmark data

From the MCEC algorithm, for each dataset, three values for the EC time point (n) were extracted. The first value is obtained from the difference in entropy measure: n such that $H(C|A_n) - H(C|A_nB_n) = 0.3 \times [H(C|A_1) - H(C|A_1B_1)]$, which means that n corresponds to the time point where a reduction of 70% from the initial value of entropy is verified, henceforth called $CH - 70$. The second and third values are a result of the minimization of $MDL(D|\mathscr{S}_n)$ and $AIC(D|\mathscr{S}_n)$, respectively, i.e. n consists of the time point where the criteria is minimum. A percentage value is associated with the EC time point:

$$\text{Earliness}[\%] = \frac{n}{L} \times 100. \quad (26)$$

This measure quantifies the amount of the TS considered necessary for a satisfactory prediction, with respect to its total length (L). The lower the value of Earliness, the less time points are considered required, and the earlier the classification is expected to be performed.

The data classification was performed with seven classifiers (Table 1). The classifier with the highest accuracy was selected. The three measures from the MCEC algorithm determine the instant from which the information in the TS can be neglected. Based on the three values of n, the selected classifier was used for the classification of the data. At most, three derivative subsets were considered, each with L defined as one of the EC time points computed by the proposed method. The preprocessing of the data included the aggregation of both training and test subsets in one single dataset (data integration). In addition, a supervised discretization by Fayyad & Irani's MDL method [8] was performed to the numeric attributes (data transformation). Furthermore, the cases which contained TS with different lengths (within the same dataset) were adjusted, that is, for each example, all instances were set to a value of L equal to the shortest sample length.

4.1. Univariate Time Series

The UEA & UCR Time Series Classification Repository [2] provides more than 90 TS datasets for research into TS classification. For analysing the performance of the MCEC algorithm, 20 benchmark datasets from the referred repository were tested. This subset of examples is considered comprehensive and representative, since it comprises a diverse range of both dimensional parameters and classification conditions. Each dataset is composed of numeric UTS ($N = 1$) with a fixed length, and their respective class labels.

The results from Table 2 describe the MCEC algorithm effort in attempting EC, based on the analysis of the information contained in the datasets. For each dataset, the first row indicates the EC time point (n); the second and third include the Earliness and Accuracy percentages, respectively; and the last row denotes the classifier with the best accuracy for the given data. The column "Full" contains the outcomes for the complete TS and it is used as a reference framework. Moreover, the "MCEC algorithm" columns indicate the results for the incomplete TS, where L is defined according to the

Dataset	MCEC algorithm			Full
	$CH-70$	MDL	AIC	
Adiac	14	1	1	–
37 classes	7.95%	0.57%	0.57%	–
$L = 176$	41.23%	17.54%	17.54%	77.47%
$w = 781$	SMO	SMO*	SMO*	SMO
ArrowHead	37	1	4	–
3 classes	14.74%	0.40%	1.59%	–
$L = 251$	68.25%	53.56%	56.87%	93.37%
$w = 211$	RandFor	RandFor*	kNN	RandFor
Beef	118	1	5	–
5 classes	25.11%	0.21%	1.06%	–
$L = 470$	60.00%	40.00%	48.33%	75.00%
$w = 60$	kNN*	kNN*	kNN*	kNN*
BeetleFly	431	107	333	–
2 classes	84.18%	20.90%	65.04%	–
$L = 512$	85.00%	67.50%	87.50%	95.00%
$w = 40$	RandFor*	NB*	RandFor*	RandFor*
BirdChicken	267	201	202	–
2 classes	52.15%	39.26%	39.45%	–
$L = 512$	77.50%	70.00%	75.00%	90.00%
$w = 40$	RandFor	NB*	NB*	NB*
Car	127	1	27	–
4 classes	22.01%	0.17%	4.68%	–
$L = 577$	69.17%	34.17%	42.50%	83.33%
$w = 120$	RandFor	kNN*	kNN*	kNN
CBF	8	1	3	–
3 classes	6.25%	0.78%	2.34%	–
$L = 128$	52.15%	44.84%	48.39%	99.68%
$w = 930$	SMO	NB*	NB*	SMO
ChlorineConc	48	1	38	–
3 classes	28.92%	0.60%	22.89%	–
$L = 166$	82.05%	54.89%	74.86%	98.98%
$w = 4307$	RandFor	RandFor*	RandFor	RandFor
Coffee	43	23	26	–
2 classes	15.04%	8.04%	9.09%	–
$L = 286$	89.29%	76.79%	80.36%	100.00%
$w = 56$	RandFor	RandFor*	NB*	RandFor*
Computers	303	1	2	–
2 classes	42.08%	0.14%	0.28%	–
$L = 720$	67.60%	58.20%	65.80%	66.00%
$w = 500$	RandFor	RandFor*	RandFor*	RandFor
Earthquakes	18	2	4	–
2 classes	3.52%	0.39%	0.78%	–
$L = 512$	89.93%	79.14%	79.86%	99.28%
$w = 278$	RandFor	RandFor*	NB*	RandFor
ECG200	16	3	6	–
2 classes	16.67%	3.13%	6.25%	–
$L = 96$	81.00%	66.50%	81.00%	90.50%
$w = 200$	BN	kNN*	SMO	kNN
FiftyWords	38	7	7	–
50 classes	14.07%	2.59%	2.59%	–
$L = 270$	32.16%	15.69%	15.69%	67.62%
$w = 905$	SMO	SMO*	SMO*	SMO
GunPoint	36	1	23	–
2 classes	24.00%	0.67%	15.33%	–
$L = 150$	92.00%	71.50%	83.00%	99.50%
$w = 200$	SMO*	SMO*	RandFor	SMO
Meat	72	1	11	–
3 classes	16.07%	0.22%	2.46%	–
$L = 448$	90.00%	66.67%	75.00%	100.00%
$w = 120$	REPTree	RandFor	SMO*	SMO*
OliveOil	55	3	6	–
4 classes	9.65%	0.53%	1.05%	–
$L = 570$	68.33%	55.00%	58.33%	96.67%
$w = 60$	RandFor	SMO*	NB*	SMO*
SwedishLeaf	7	1	2	–
15 classes	5.47%	0.78%	1.56%	–
$L = 128$	56.53%	29.60%	37.78%	91.02%
$w = 1125$	RandFor	SMO*	SMO*	SMO
SynthControl	5	1	2	–
6 classes	8.33%	1.67%	3.33%	–
$L = 6$	82.17%	49.67%	69.00%	98.83%
$w = 600$	BN	BN*	BN	BN
TwoPatterns	95	1	11	–
4 classes	74.22%	0.78%	8.59%	–
$L = 128$	56.36%	26.28%	29.20%	75.18%
$w = 5000$	RandFor	RandFor*	kNN	RandFor
Wafer	11	2	3	–
2 classes	7.24%	1.32%	1.97%	–
$L = 152$	97.91%	97.60%	97.66%	99.85%
$w = 7164$	kNN	RandFor*	kNN	RandFor

Table 2: MCEC algorithm experimental results on UTS.

values of n. The symbol (*) means that more than one classifier achieved the best accuracy.

The value of Earliness is always beneath 100% for all the measures of the MCEC algorithm. Concerning $CH-70$, the accuracy with less time points outperforms the reference value ("Full" column) only for the "Computers" dataset. This example suggests that it is possible to obtain a better classification performance using only part of the TS from the data. In the "Wafer" outcomes, using only 7.24% of the TS, an accuracy of 97.91% is achieved, which consists of -1.94% in comparison with the full-length result. This means that, in these experiments, with fewer time points (earlier in time), the loss in terms of classification accuracy can be low. Concerning the 18 datasets with $n_{MDL} \neq n_{AIC}$, the classification accuracy results of AIC outperform the ones for MDL in all cases. This suggests that, based on these experiments, AIC surpasses MDL, with respect to accuracy. However, in all situations, the accuracy for both criteria is lower than for the full-length data. In addition, the results show that $n_{MDL} = 1$ for 12 of the 20 cases, and $n_{AIC} = 1$ for only 1 dataset ("Adiac").

From the comparison between the three measures, $CH - 70$ achieves higher classification accuracy in 18 of the 20 cases, AIC in 1 dataset ("BeetleFly"), and in 1 example ("ECG200") a draw is verified between the difference in entropy and AIC. MDL has a percentage of correctly classified instances always lower or equal than the other measures. Regarding the Earliness percentage, except for the events where $n_{MDL} = n_{AIC}$, MDL proposes always the lowest values for the EC time point. Therefore, in general, $CH - 70$ achieves better results, in terms of accuracy, and MDL demonstrates a superior earliness ability. AIC evidences the foremost competence in balancing these two targets. Nevertheless, the EC capabilities of the MCEC algorithm are acknowledged, seeing that this context is based on the tradeoff between both objectives: accuracy and earliness.

4.2. Multivariate Time Series

As a supplement to a study on MTS Classification [3], a group of investigators gathered a collection of datasets, useful for experiments on methods that deal with this type of data. These examples were obtained from a variety of sources, such as repositories [20, 6] and other websites. For analysing the performance of the proposed algorithm, six benchmark datasets were selected from the available resources, as an attempt to provide experimental results in an expansive set of conditions. Table 3 lists the results of the experiments performed. The process used for the UTS was replicated in these six experiments.

Regarding both model selection criteria (columns MDL and AIC), Earliness $< 100\%$ for all datasets. The proposed EC time points from both scoring functions are coincident in 4 of the 6 cases. In all these examples, where $n_{MDL} = n_{AIC}$, the criteria sugges-

Dataset	MCEC algorithm			Full
	$CH-70$	MDL	AIC	
ECG ($N=2$)	13	1	3	–
2 classes	33.33%	2.56%	7.69%	–
$L=39$	87.00%	77.00	80.00%	86.00%
$w=200$	RandFor	SMO*	J48	SMO
JapanVow ($N=12$)	2	1	1	–
9 classes	28.57%	14.29%	14.29%	–
$L=7$	88.13%	85.47%	85.47%	94.53%
$w=640$	RandFor	SMO	SMO	SMO
Libras ($N=2$)	14	1	1	–
15 classes	31.11%	2.22%	2.22%	–
$L=45$	60.28%	30.28%	30.28%	79.17%
$w=360$	kNN	RandFor	RandFor	RandFor
PenDigits ($N=2$)	3	1	1	–
10 classes	37.50%	12.50%	12.50%	–
$L=8$	78.26%	47.95%	47.46%	98.45%
$w=10992$	RandFor	RandFor	RandFor	SMO
RobotLP1 ($N=6$)	2	1	1	–
4 classes	13.33%	6.67%	6.67%	–
$L=15$	89.77%	84.09%	82.96%	95.46%
$w=88$	kNN	NB*	NB*	SMO
Wafer ($N=6$)	51	1	24	–
2 classes	49.04%	0.96%	23.08%	–
$L=104$	95.31%	90.29%	93.55%	98.49%
$w=1194$	SMO*	SMO*	RandFor	SMO

Table 3: MCEC algorithm experimental results on MTS.

tion consists of using only the first instant of the TS for classifying the data. In 2 of those 4 cases, the classification accuracy is above 80% ("JapaneseVowels" and "RobotLP1"), which corresponds to a considerably decent outcome. Table 3 shows $n_{AIC}=1$ in 4 of the 6 cases, and $n_{MDL}=1$ for all experiments.

With regard to the examples where $n_{MDL} \neq n_{AIC}$ ("ECG" and "Wafer"), $n_{MDL} < n_{AIC}$ for all cases. However, seeing that the largest accuracy difference between MDL and AIC is 3.26% ("Wafer"), these results suggest that by giving priority to earliness, the percentage of correctly classified instances is not extensively affected. In fact, for both model selection criteria, the values from n obtained Accuracy $\geq 70\%$ in 4 of the 6 cases, which assigns some confidence to the MCEC algorithm in analysing the EC opportunity. When comparing the three measures, the difference in entropy achieves higher classification accuracy in all cases. Concerning Earliness, the model selection criteria obtain always the lowest values, and, particularly, MDL achieves the best results. In general, the difference in entropy measure performs better with regard to classification accuracy and MDL manifests a higher disposition to earliness. However, the most efficient tradeoff between these two requirements seems to be found for AIC. These conclusions are in line with the inferences drawn from the experiments with UTS.

4.3. Wilcoxon signed-ranks sum test

The univariate and multivariate experimental results were compared with statistical significance tests in order to understand the benefit of the tradeoff between the two main goals in EC: accuracy and earliness. Among the tested datasets, the MCEC algorithm provided a value of n, with an associated percentage (Earliness). For each situation, the group of classifiers determined the Accuracy value. In addition, the classification of the full-data worked as a reference framework: no earliness and complete TS accuracy. Aiming for a representation of the balance between these two requirements, a mathematical expression can be defined as:

$$BEA(p) = p \times (100 - E) + (1 - p) \times A, \qquad (27)$$

where E and A correspond to the Earliness and Accuracy percentages, respectively; and p consists of the weight that determines the relevance given to each variable. Seeing that an accurate classification is desirable, as early as possible, Eq. (27) describes the management of the two fundamental challenges of the EC problem. The 26 datasets from Tables 2 and 3 were considered, as well as their respective values of E and A, for each of the three measures that compose the MCEC algorithm, together with the reference framework. Note that all Full outcomes verify $E = 100\%$, since the entire TS are considered for classification.

Table 4 includes the results of the Wilcoxon signed-rank sum test [27], for comparing the performance of the MCEC algorithm measures. These tests examine

Comparison		p = 0	0.25	0.5	0.75
$CH-70 \Leftrightarrow MDL$	size	26	26	26	26
	p-value	<0.01	<0.01	0.81	<0.01
	better	$\Leftarrow$	$\Leftarrow$	$\rightarrow$	$\Rightarrow$
$CH-70 \Leftrightarrow AIC$	size	25	26	26	26
	p-value	<0.01	0.01	0.34	<0.01
	better	$\Leftarrow$	$\Leftarrow$	$\rightarrow$	$\Rightarrow$
$CH-70 \Leftrightarrow$ Full	size	26	26	26	26
	p-value	<0.01	<0.01	<0.01	<0.01
	better	$\Rightarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$
$MDL \Leftrightarrow AIC$	size	20	20	20	20
	p-value	<0.01	<0.01	0.23	0.37
	better	$\Rightarrow$	$\Rightarrow$	$\rightarrow$	$\leftarrow$
$MDL \Leftrightarrow$ Full	size	26	26	26	26
	p-value	<0.01	0.70	<0.01	<0.01
	better	$\Rightarrow$	$\rightarrow$	$\Leftarrow$	$\Leftarrow$
$AIC \Leftrightarrow$ Full	size	26	26	26	26
	p-value	<0.01	0.37	<0.01	<0.01
	better	$\Rightarrow$	$\leftarrow$	$\Leftarrow$	$\Leftarrow$

Table 4: Comparison of the MCEC algorithm measures and Full against each other, using the Wilcoxon signed-rank sum test applied to the tradeoff experimental data, scored according to $BEA(p)$.

the relation between the measures in pairs in order to verify if there is enough evidence to claim that the differences are significant, for a significance level of $\alpha = 0.05$. The arrow in Table 4 points towards the measure with better performance, according to the value of $p \in \{0, 0.25, 0.5, 0.75\}$. Double arrow means there is enough evidence to claim the difference is significant.

The results demonstrate that, for $p = 0$, there is enough evidence to claim that Full surpasses all other measures. Furthermore, between $CH-70$ and the model selection criteria, the difference in entropy outperforms both scoring functions, and AIC shows better results than MDL. All these differences are statistically significant. For $p = 0.25$, $CH-70$ has the best performance in comparison with all the remaining. The AIC

measure seems to achieve significantly superior results than MDL, however, there is not enough evidence to claim that AIC outperforms Full, nor that the latter surpasses MDL. For $p = 0.5$, the only assurance consists of Full performing the worst. Among $CH - 70$, MDL and AIC, the differences between them are not statistically significant. Lastly, at $p = 0.75$, Full continues to be surpassed by all the others, as well as the difference in entropy in comparison with both model selection criteria. However, between MDL and AIC, there is not enough evidence to confirm which performs the best.

5. Conclusions

The achieved outcomes confirm the ability of the MCEC method to examine the EC opportunity within a dataset. In general, the three main measures are capable of choosing an early time point based on which the TS classification is plausible. Overall, the first measure obtains better accuracy results, MDL demonstrates a superior tendency for earliness, and AIC attains the most competent balance between both aims. While AIC is known to select a model more readily, MDL's choice is considered more consistent. The large number of examples where $n_{MDL} = 1$ may indicate that, given the information available, the criterion recognized that the increase in the knowledge obtained from the data did not justify the growth in the model complexity required for describing it. Conversely, the AIC results demonstrate a more adventurous disposition in choosing the value for n, and, in these experiments, that seems to have produced relative success. Although AIC seems to surpass MDL for $p = 0.5$, and the latter appears to outperform the difference in entropy, these inferences are not statistically significant at the $\alpha = 0.05$ significance level. On the other hand, the difference in entropy is surpassed by both model selection criteria, for $p = 0.75$. However, in this case, there is not enough statistical evidence to claim that MDL outperforms AIC, in spite of the empirical outcomes. Conversely, for $p = 0.25$, the entropy measure surpasses both scoring functions, and AIC obtains better results than MDL. Herein, these comparisons are statistically significant.

The MCEC algorithm can be extended to deal with datasets where the TS length and the number of attributes per time point vary among all instances. In addition, a classification method can be developed based on the capabilities of this information-theoretic approach. In this case, the algorithm would be able to assign a class label to a new single incomplete TS. Finally, the feature selection potentialities of the MCEC method can be exploited. In particular, a greedy feature selection could be performed based, not only on the difference in entropy measure, but also on the model selection criteria.

Acknowledgements

The author would like to thank to Prof. Alexandra Carvalho, Prof. Paulo Mateus, Prof. Dr. Helena Canhão, Cátia Botas and Mariano Lemus. A special acknowledgement to IT for the possibility of having a grant from NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014).

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans Automat Contr*, 19(6):716–723, 1974.

[2] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 2016.

[3] M. G. Baydogan and G. Runger. Learning a symbolic representation for multivariate time series classification. *Data Min. Knowl. Discov.*, 29(2):400–422, Mar 2015.

[4] I. Ben-Gal. Bayesian networks. *Encyclopedia of statistics in quality and reliability*, 2008.

[5] A. M. Carvalho. Scoring functions for learning Bayesian networks. *INESC-ID Tec. Rep*, 2009.

[6] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[8] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI-93, 1993*, pages 1022–1029, 1993.

[9] M. F. Ghalwash and Z. Obradovic. Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC Bioinformatics*, 13:195, 2012.

[10] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *ICDM'13*, pages 201–210, 2013.

[11] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Utilizing temporal patterns for estimating uncertainty in interpretable early decision making. In *SIGKDD'14*, pages 402–411, 2014.

[12] M. P. Griffin and J. R. Moorman. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*, 107(1):97–104, 2001.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[14] J. Hamilton. *Time series analysis*. Princeton Univ. Press, Princeton, NJ, 1994.

[15] N. Hatami and C. Chira. Classifiers with a reject option for early time-series classification. *CoRR*, abs/1312.3989, 2013.

[16] G. He, Y. Duan, R. Peng, X. Jing, T. Qian, and L. Wang. Early classification on multivariate time series. *Neurocomputing*, 149:777–787, 2015.

[17] G. He, Y. Duan, T. Qian, and X. Chen. Early prediction on imbalanced multivariate time series. In *CIKM'13*, pages 1889–1892, 2013.

[18] D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[19] K. Li, S. Li, and Y. Fu. Early classification of ongoing observation. In *ICDM'14*, pages 310–319, 2014.

[20] M. Lichman. UCI machine learning repository, 2013. `http://archive.ics.uci.edu/ml/`.

[21] Y. Lin, H. Chen, V. S. Tseng, and J. Pei. Reliable early classification on multivariate time series with numerical and categorical attributes. In *PAKDD'15*, pages 199–211, 2015.

[22] U. Mori, A. Mendiburu, E. J. Keogh, and J. A. Lozano. Reliable early classification of time series based on discriminating the classes over time. *Data Min. Knowl. Discov.*, 31(1):233–263, 2017.

[23] A. Mueen, E. J. Keogh, and N. E. Young. Logical-shapelets: an expressive primitive for time series classification. In *SIGKDD'11*, pages 1154–1162, 2011.

[24] N. Parrish, H. S. Anderson, M. R. Gupta, and D. Hsiao. Classifying with confidence from incomplete information. *J. Mach. Learn. Res.*, 14(1):3561–3589, 2013.

[25] J. J. Rodrıguez and C. J. Alonso. Boosting interval-based literals: Variable length and early classification. *Knowledge Discovery from Temporal and Spatial Data (W12)*, 2002.

[26] W. Wang, C. Chen, W. Wang, P. Rai, and L. Carin. Earliness-aware deep convolutional networks for early time series classification. *CoRR*, abs/1611.04578, 2016.

[27] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[28] Z. Xing, J. Pei, G. Dong, and P. S. Yu. Mining sequence classifiers for early prediction. In *SDM'08*, pages 644–655, 2008.

[29] Z. Xing, J. Pei, and P. S. Yu. Early classification on time series. *Knowl. Inf. Syst.*, 31(1):105–127, 2012.

[30] Z. Xing, J. Pei, P. S. Yu, and K. Wang. Extracting interpretable features for early classification on time series. In *SDM'11*, pages 247–258, 2011.

[31] L. Ye and E. J. Keogh. Time series shapelets: a new primitive for data mining. In *SIGKDD'09*, pages 947–956, 2009.