Appendix A

Synthetic example of the proposed method

This appendix describes the implementation details of the proposed method through the example of a synthetically generated dataset. As previously mentioned, for a given multivariate time series $T_i = (X_1, X_2, ..., X_L)$, where L corresponds to its time length, each component $X_k = (X_{k_1}, X_{k_2}, ..., X_{k_N})$ consists of a set of N features measured at time point k. A class label c_i is associated to each T_i through the relation $Class(T_i) = c_i$, and the dataset D to be analysed consists of a collection of pairs $(T_i, c_i) : i \in \{1, ..., w\}$, where w corresponds to the number of instances.

Consider the example represented in Table A.1. This dataset consists of ten instances $(w = 10)$ of univariate time series $(N = 1)$ with seven time points each $(L = 7)$. The feature described over time is of type boolean $(X_k \in \{0, 1\})$, and the class label attribute includes two alternatives $(c_i \in \{\text{C}0, \text{C}1\})$. The first two time points $(X_1$ and $X_2)$ are randomly generated and all the others correspond to the

Table A.1: Synthetic dataset example.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	c_i
0	1	1	0	1	1	0	C1
1	1	0	1	1	0	1	C1
0	1	1	0	1	1	0	C1
1	1	0	1	1	0	1	C1
1	0	1	1	0	1	1	C0
1	0	1	1	0	1	1	C0
1	1	0	1	1	0	1	C1
0	1	1	0	1	1	0	C1
1	0	1	1	0	1	1	C0
0	1	1	0	1	1	0	C1

XOR (Table A.2) of the two previous time points. For example, in the first instance, X_3 corresponds to $X_1 \oplus X_2 = 1$, and X_4 is equal to $X_2 \oplus X_3 = 0$. The class label c_i is the result of the exclusive disjunction between X_6 and X_7, concatenated with the letter "C".

Table A.2: Exclusive disjunction (XOR).

x	y	$x \oplus y$
0	0	0
0	1	1
1	0	1
1	1	0

According to the proposed approach, the multivariate time series are decomposed in

$$T_i = (X_1, X_2, ..., X_n, X_{n+1}, ..., X_{L-1}, X_L)\,, \tag{A.1}$$

and the dataset is organized in three groups:

$$\begin{aligned} A_n &= \{X_1, X_2, ..., X_n\}\,, \\ B_n &= \{X_{n+1}, X_{n+2}, ..., X_L\}\,, \\ C &= \{c_i\}\,. \end{aligned} \tag{A.2}$$

A.1 Difference in entropy

Once aiming for early classification, we are interested in predicting the class label of a time series as early as possible, provided that the classification accuracy is close to the one using the complete data. For this purpose, in the first stage, the conditional entropy of the dataset is the subject of study. The goal consists of analysing the difference in entropy:

$$H(C|A_n) - H(C|A_nB_n), \tag{A.3}$$

while varying the early classification time point n from $\{1, ..., L\}$.

Considering the three groups described in Equation (A.2) as well as the definition from Equation (3.3), the calculation of the conditional entropies is performed through:

$$\begin{aligned} H(C|A_n) &= \sum_{a,c} p(A_n = a, C = c) \log_2 \left[\frac{p(A_n = a)}{p(A_n = a, C = c)} \right], \\ H(C|A_nB_n) &= \sum_{a,b,c} p(A_n = a, B_n = b, C = c) \log_2 \left[\frac{p(A_n = a, B_n = b)}{p(A_n = a, B_n = b, C = c)} \right]. \end{aligned} \tag{A.4}$$

The conditional entropy $H(C|A_nB_n)$ quantifies the amount of information needed to describe the outcome of the class label c_i, based on the knowledge of the entire time series. This value is constant when varying n from $\{1, ..., L\}$, since it represents the lowest possible uncertainty in the outcome of C. On the other hand, $H(C|A_n)$ quantifies the amount of information needed to predict c_i, provided that there is only information until time point n. This value is expected to decrease with the increase of n, since the growth on the amount of available information is expected to reduce the uncertainty of the prediction.

The statistical parameters included in Equation (A.4) are estimated as the quotient between the

number of occurrences of each specific case and the total number of instances in the dataset:

$$\begin{aligned} p(A_n = a) &= \frac{\text{number of occurrences of } \{a\}}{w}, \\ p(A_n = a, B_n = b) &= \frac{\text{number of occurrences of } \{a, b\}}{w}, \\ p(A_n = a, C = c) &= \frac{\text{number of occurrences of } \{a, c\}}{w}, \\ p(A_n = a, B_n = b, C = c) &= \frac{\text{number of occurrences of } \{a, b, c\}}{w}, \end{aligned} \tag{A.5}$$

for which $\{a\}$, $\{a, b\}$, $\{a, c\}$ and $\{a, b, c\}$ represent the existing cases included in the respective group described in Equation (A.2).

For $n = 1$, the organization of the time series in three groups corresponds to:

$$A_1 = \{X_1\}, \quad B_1 = \{X_2, X_3, X_4, X_5, X_6, X_7\}, \quad C = \{c_i\}, \tag{A.6}$$

and the information from the dataset can be structured in lists such as:

$$\begin{aligned} \mathcal{A}_1 &= \begin{bmatrix} (\{0\}, 4) \\ (\{1\}, 6) \end{bmatrix}, \quad \mathcal{AC}_1 = \begin{bmatrix} (\{1\,\mathsf{C0}\}, 3) \\ (\{1\,\mathsf{C1}\}, 3) \\ (\{0\,\mathsf{C1}\}, 4) \end{bmatrix}, \\ \mathcal{AB} &= \begin{bmatrix} (\{1011011\}, 3) \\ (\{1101101\}, 3) \\ (\{0110110\}, 4) \end{bmatrix}, \quad \mathcal{ABC} = \begin{bmatrix} (\{1011011\,\mathsf{C0}\}, 3) \\ (\{0110110\,\mathsf{C1}\}, 4) \\ (\{1101101\,\mathsf{C1}\}, 3) \end{bmatrix}. \end{aligned} \tag{A.7}$$

where the format consists of $\mathcal{G} = [(\{g\}, \text{number of occurrences})]$. Considering the list $\mathcal{A}_1$, the observation $\{0\}$ for X_1 occurs 4 times, and $X_1 = 1$ is verified in 6 of the 10 instances from the data. From these lists, the parameters described in Equation (A.5) are calculated as:

$$\begin{aligned} p(A_1 = \{0\}) = \frac{4}{10} = 0.4 \text{ bits}, \quad p(A_1 = \{1\}) = \frac{6}{10} = 0.6 \text{ bits}, \\ p(A_1C = \{1\,\mathsf{C0}\}) = p(A_1C = \{1\,\mathsf{C1}\}) = 0.3 \text{ bits}, \quad p(A_1C = \{0\,\mathsf{C1}\}) = 0.4 \text{ bits}, \\ p(A_1B_1 = \{1011011\}) = p(A_1B_1 = \{1101101\}) = \frac{3}{10} = 0.3 \text{ bits}, \\ p(A_1B_1 = \{0110110\}) = 0.4 \text{ bits}, \\ p(A_1B_1C = \{1011011\,\mathsf{C0}\}) = p(A_1B_1C = \{1101101\,\mathsf{C1}\}) = 0.3 \text{ bits}, \\ p(A_1B_1C = \{0110110\,\mathsf{C1}\}) = 0.4 \text{ bits}, \end{aligned} \tag{A.8}$$

and the conditional entropies described in Equation (A.4) are computed through:

$$\begin{aligned} H(C|A_1) &= \sum_{a,c} p(A_1C = \{a, c\}) \log_2 \left[\frac{p(A_1 = \{a\})}{p(A_1C = \{a, c\})} \right] = \\ &= 0.3 \log_2 \left[\frac{0.6}{0.3} \right] + 0.3 \log_2 \left[\frac{0.6}{0.3} \right] + 0.4 \log \left[\frac{0.4}{0.4} \right] = 0.6 \text{ bits}; \end{aligned} \tag{A.9}$$

$$H(C|A_1B_1) = \sum_{a,b,c} p(A_1B_1C = \{a,b,c\}) \log_2 \left[\frac{p(A_1B_1 = \{a,b\})}{p(A_1B_1C = \{a,b,c\})} \right] = \\ = 2 \left(0.3 \log_2 \left[\frac{0.3}{0.3} \right] \right) + 0.4 \log_2 \left[\frac{0.4}{0.4} \right] = 0. \tag{A.10}$$

On the one hand, $H(C|A_1) = 0.6$ bits represents the amount of information needed to predict the classes of the time series, given that X_1 is known. On the other hand, $H(C|A_1B_1) = 0$ indicates that the complete time series provide enough information for describing the group C. The difference in entropy, equal to $H(C|A_1) - H(C|A_1B_1) = 0.6$ bits, denotes that with only the first time point of the time series there is still a lack of information for predicting the class labels.

For $n = 2$, the organization of the time series in three groups becomes:

$$A_2 = \{X_1, X_2\}, \quad B_2 = \{X_3, X_4, X_5, X_6, X_7\}, \quad C = \{c_i\}, \tag{A.11}$$

and the data structured in lists correspond to:

$$\mathcal{A}_2 = \begin{bmatrix} (\{10\}, 3) \\ (\{11\}, 3) \\ (\{01\}, 4) \end{bmatrix}, \quad \mathcal{AC}_2 = \begin{bmatrix} (\{11\,\mathsf{C1}\}, 3) \\ (\{10\,\mathsf{C0}\}, 3) \\ (\{01\,\mathsf{C1}\}, 4) \end{bmatrix}. \tag{A.12}$$

Note that the lists $\mathcal{AB}$ and $\mathcal{ABC}$ are the same as in Equation (A.7) since they do not change with the variation of n. In this case, the computation of the difference in entropy is equal to:

$$H(C|A_2) - H(C|A_2B_2) = 0 - 0 = 0. \tag{A.13}$$

This result denotes that with the first two time points of the time series (X_1 and X_2) there is enough information for predicting the class labels.

Figure A.1 describes the evolution of the difference in entropy from Equation (A.3) for $n \in \{1, ..., 10\}$. Since $H(C|A_n) - H(C|A_nB_n) = 0$ for $n \geq 2$, the correlations between the early states of T_i and the classes c_i are completely represented using only the first two time points. It is possible to infer that the information given by the time series after X_2 does not provide any useful knowledge about the class label attribute.

A.2 Complexity of the model

In the second stage, based on two Bayesian network scoring functions, the complexity of the model is examined in the interest of choosing the early time point, which is able not only to achieve an early classification, but also to consider the simplicity of the choice. The goal consists of analysing the function:

$$\phi(\mathscr{S}_n|D) = \alpha \cdot |\mathscr{S}_n| - LL(\mathscr{S}_n|D), \tag{A.14}$$

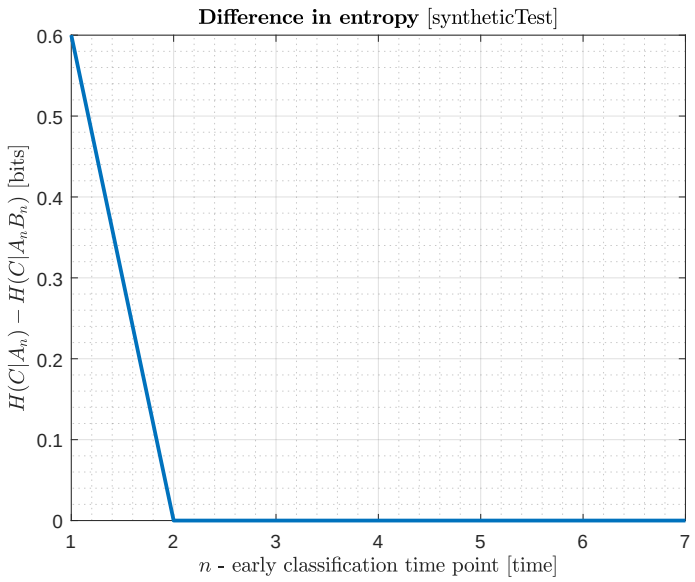


Figure A.1: Variation of the entropy difference while $n \in \{1, ..., L\}$, for the data represented in Table A.1.

while varying the early classification time point n from $\{1, ..., L\}$, where $|\mathscr{S}_n|$ is described in Equation (3.22) and $LL(\mathscr{S}_n|D)$ is defined as:

$$\begin{aligned} LL(\mathscr{S}_n|D) &= \sum_{i=1}^{w} \log_2 \left[p(C = c|A_n = a)p(B_n = b|A_n = a)p(A_n = a)\right] = \\ &= \sum_{i=1}^{w} \log_2 \left[\frac{p(A_n = a, B_n = b)p(A_n = a, C = c)}{p(A_n = a)}\right], \end{aligned} \tag{A.15}$$

according to the Bayes' theorem. The value of α is independent from the early time point, and while for the MDL score, $\alpha = \frac{1}{2}\log_2 w$, for the AIC, $\alpha = 1$. The statistical parameters for computing the log-likelihood of the model given the data are estimated as described in Equation (A.5).

Regarding the number of independent parameters, $||C||$ denotes the number of distinct cases in group C, i.e. the number of classes in the dataset. This value is constant while varying n, seeing that the variation of the early time point does not affect group C. Similarly, $||A_n||$ corresponds to the number of different existing cases in group A_n. This value is expected to increase with n, as a greater amount of analysed instants from the time series leads to a higher number of possible cases. At some point, $||A_n||$ is expected to stabilize, possibly when the information added is redundant, and consequently, unnecessary for the prediction.

For $n = 1$, the groups are organized as represented in Equation (A.6) and the information contained in the dataset can be structured in lists such as the ones denoted in Equation (A.7). From the statistical

parameters calculated in Equation (A.8), the log-likelihood can be computed as:

$$\begin{aligned} -LL(\mathscr{S}_1|D) &= -\sum_{i=1}^{10} \log_2 \left[\frac{p(A_1B_1 = \{a,b\}) \cdot p(A_1C = \{a,c\})}{p(A_1 = \{a\})} \right] = \\ &= 3\log_2 \left[\frac{0.3 \cdot 0.3}{0.6} \right] + 4\log_2 \left[\frac{0.4 \cdot 0.4}{0.4} \right] + 3\log_2 \left[\frac{0.3 \cdot 0.3}{0.6} \right] = \\ &= 21.7095 \text{ bits.} \end{aligned} \tag{A.16}$$

Note that this value corresponds to the amount of information required to represent the dataset D using the model $\mathscr{S}_1$, and the sum comprises all the instances included in D (in this case, $w = 10$). Since $||A_1||$ consists of the number of distinct cases in group $\mathcal{A}_1$, and $||C||$ denotes the number of classes in the dataset, from Equation (3.22), the number of independent parameters in the model is equal to:

$$|\mathscr{S}_1| = ||A_1|| \cdot ||C|| - 1 = 2 \cdot 2 - 1 = 3 \text{ bits.} \tag{A.17}$$

This value quantifies the amount of information needed to encode the model $\mathscr{S}_1$, as well as the data D given the model. It can be viewed as a measure of the complexity associated to using the model $\mathscr{S}_1$ to represent the dataset from Table A.1.

Concerning the AIC score, seeing that $\alpha = 1$ and according to Equation (3.21), its computation corresponds to:

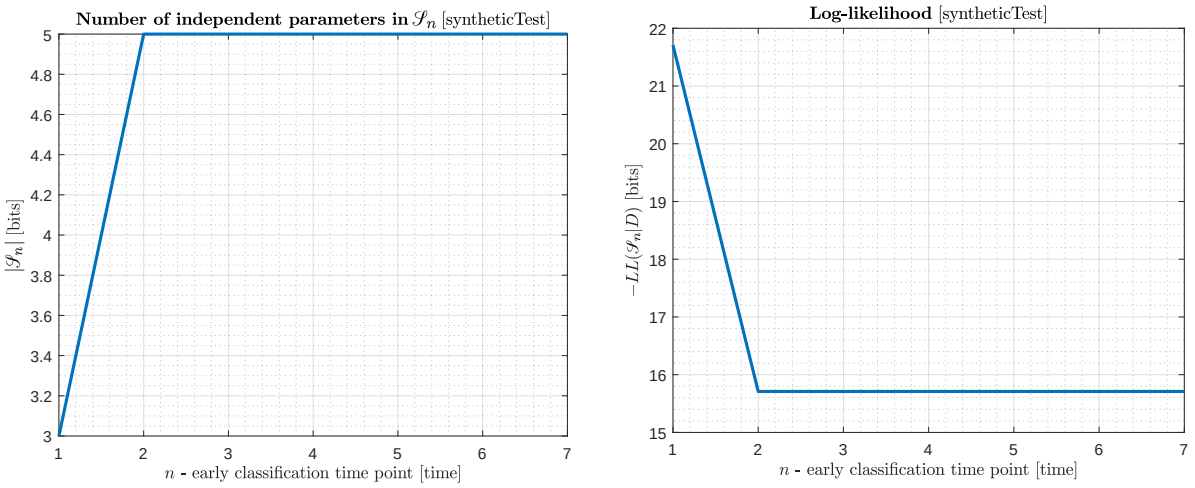
$$AIC(\mathscr{S}_1|D) = |\mathscr{S}_1| - LL(\mathscr{S}_1|D) = 3 + 21.7095 = 24.7095 \text{ bits.} \tag{A.18}$$

With regard to the MDL scoring function, the penalization factor is $\alpha = \frac{1}{2}\log_2 10 = 1.661$ bits, and through Equation (3.20) this score is calculated as:

$$MDL(\mathscr{S}_1|D) = \alpha \cdot |\mathscr{S}_1| - LL(\mathscr{S}_1|D) = 1.661 \cdot 3 + 21.7095 = 26.6925 \text{ bits.} \tag{A.19}$$

In these sort of model selection, the idea is to find the $\mathscr{S}_n$ that is good enough to capture the information in the data D, but not so complex that it makes the choice infeasible. The multiple models represent the variation of n from $\{1, ..., L\}$, which means that there are as many $\mathscr{S}_n$ as the number of time points (L). By minimizing the general function $\phi(\mathscr{S}_n|D)$ from Equation (A.14), we are trying to find a balance between the complexity of the model and its ability to fit to the data. The goal is to find the early time point for which both $MDL(\mathscr{S}_n|D)$ and $AIC(\mathscr{S}_n|D)$ are as low as possible, meaning that the information contained in the time series until n is enough to represent the dataset in an effective but simple manner.

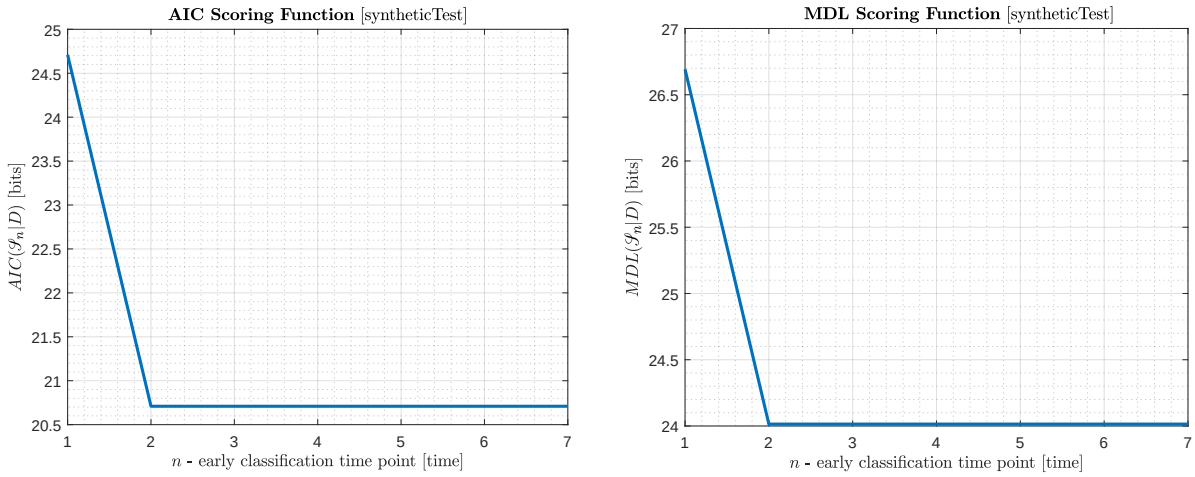
Aiming for a more detailed analysis of the terms that compose $\phi(\mathscr{S}_n|D)$, Figure A.2(a) denotes the variation of the number of independent parameters in $\mathscr{S}_n$ with n, and Figure A.2(b) represents the graph of the log-likelihood term for all time points. As depicted in Figure A.2(a), the complexity of the model increases from $n \in \{1, 2\}$, i.e. the more instants from the time series analysed, the higher the amount of information needed to encode $\mathscr{S}_n$. Notice that since $|\mathscr{S}_n|$ is constant for $n \geq 2$, the information added to the model in this interval does not affect its complexity. From Figure A.2(b), the significant decrease of the log-likelihood at $n = 2$ followed by a stabilization from that time point on indicates that the dataset



(a) $|\mathscr{S}_n|$ (b) $-LL(\mathscr{S}_n|D)$

Figure A.2: Variation of the terms from $\phi(\mathscr{S}_n|D)$ while $n \in \{1, ..., L\}$, for the dataset from Table A.1.

D is effectively described by $\mathscr{S}_2$, i.e. using only the observations from the two initial instants of the time series. Figure A.3 represents the values for the AIC and the MDL scoring functions. Seeing that



(a) $AIC(\mathscr{S}_n|D)$ (b) $MDL(\mathscr{S}_n|D)$

Figure A.3: Variation of the scoring functions while $n \in \{1, ..., L\}$, for the dataset from Table A.1.

the difference between the scores is mainly related with the penalization factor on the complexity term, the variation is very similar in both cases, achieving a minimum value at $n \geq 2$. This denotes that a satisfactory tradeoff between the complexity of the model and its ability to represent the data is found at $n = 2$.

The same conclusion is reached for the proposed approaches. Not only from the model complexity analysis but also from the study of the difference in entropy, the results demonstrate that we are expected to be able to accurately classify the time series from the synthetic dataset, based only on the first two time points. A closer look to the two initial columns from Table A.1 (X_1 and X_2) corroborates this inference, seeing that whenever $A_2 = \{11\}$ or $A_2 = \{01\}$, the class label is C1; and whenever $A_2 = \{10\}$, the class label is C0; i.e. the knowledge of A_2 is enough to describe C with no uncertainty.

A.3 Early classification analysis

The analysis of the percentage of correctly classified instances (Equation 2.1) for the synthetic dataset is represented in Figure A.4. Except for the REPTree classifier, all the others accomplish 100% accuracy for $n \geq 2$, which matches with the conclusions obtained from the proposed method. The utility of the classification accuracy investigation is further developed in Chapter 4.

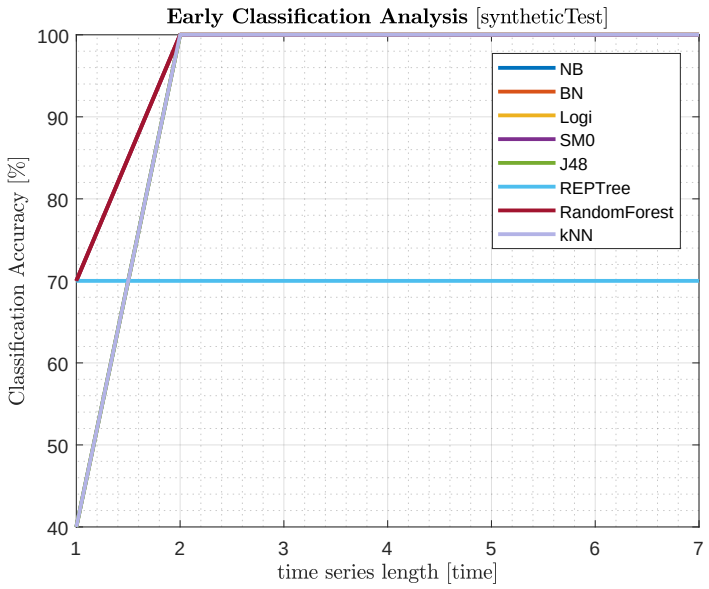


Figure A.4: Multiple classifiers performance accuracy on the data represented in Table A.1, for every time series lengths.