Mutual Information Based Feature Subset Selection in Multivariate Time Series Classification

Josu Ircio [1], Aizea Lojo [1], Usue Mori [2], and Jose A. Lozano[2,3]

[1] *Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), Arrasate-Mondragón,* {jircio, alojo}@ikerlan.es ,

[2] *University of the Basque Country UPV/EHU, 20018 San Sebastian, Spain,* {usue.mori, ja.lozano}@ehu.eus ,

[3] *Basque Center for Applied Mathematics (BCAM) 48009 Bilbao, Spain*

Abstract

This paper deals with supervised classification of multivariate time series. In particular, the goal is to propose a filter method to select a subset of time series. Consequently, we adopt the framework proposed by Brown et al. [10]. The key point in this framework is the computation of the mutual information between the features, which allows us to measure the relevance of each feature subset. In our case, where the features are a time series, we use an adaptation of existing nonparametric mutual information estimators based on the k-nearest neighbor. Specifically, for the purpose of bringing these methods to the time series scenario, we rely on the use of dynamic time warping dissimilarity. Our experimental results show that our method is able to strongly reduce the number of time series while keeping or increasing the classification accuracy.

Keywords— Multivariate time series, supervised classification, feature susbset selection, mutual information.

1 Introduction

The fourth Industrial Revolution has brought many advances in digital technologies. This global transformation has led to the development of new monitoring systems. The information reported from the monitorization could be used, for example, to detect anomalies in person vital signs or perform a predictive maintenance of a machine that prevents a breakdown. The data collected in those scenarios have, in most cases, two properties: 1) they are time ordered and 2) the relationships between closest in time data points are stronger than between the farthest data points. Therefore, they could be considered as time series (TS). Due to the ubiquity of this kind of data, in the last few decades, the development of time series specific techniques have increased [1].

One of the most common activities in time series analysis is time series classification (TSC) [2, 3]. This is a supervised learning problem, where the objective is to distinguish between two or more possible situations or classes, while taking into account the information and properties of the time series. Especially for the univariate case, where a classifier is trained using one dimension time series, a variety of methods have been proposed [4, 5, 6]. However, due to the complexity of high-dimensional data, classification of multivariate time series (MTS) has received less attention.
In comparison with the univariate case, where an instance is composed by a unique TS, in MTS classification, two or more time series represent an instance, so they need to be analyzed together to obtain a classifier [7]. This makes the problem of MTS classification more challenging.

Some MTS classification problems involving the processing of these large volumes of TS data require too many resources and can become unsustainable. Moreover, it is common that many of the collected series are redundant or there may be series that for the purpose of classification are not useful and generate noise, which penalizes the performance of the classifiers. Therefore, Feature Subset Selection (FSS) methods are a necessary pre-processing step for dealing with high-dimensional MTS classification problems [8]. The features to be selected in a MTS classification problem are univariate TS. The goals of these techniques are to avoid over-fitting, to produce easily interpretable models and to improve the classification [9].

The objective of this paper is to develop a time series subset selection method with the following properties:

- The output of the method is a subset of the original time series (not a transformation of them).
- The method takes into account the temporal information contained in the series.
- The method considers the information provided by the class variable in the selection process.

Consequently, we adapt FSS methods based on information theory designed for non-temporal data [10, 9, 11] to the multivariate time series classification scenario. We particularly concentrate on the method proposed by Brown et al. [10]. The main adaptation consists in the computation of the mutual information (MI) between two time series, and between a time series and the class variable. This computation is carried out by modifying non-parametric MI estimation methods by allowing them to account for the temporal information of the time series. The results obtained suggest that our FSS method succeeds in improving the accuracy of the MTS classification problem by reducing the univariate time series that compose the MTS.

The rest of this paper is structured as follows. In Section 2, related work of FSS techniques in time series is reviewed. Section 3 details the proposed approach for TS subset selection in MTS classification. In Section 4, adapted methods for estimating MI are described. In Section 5, the experimental framework is introduced, followed by the obtained results and discussion in Section 6. Finally, in Section 7, conclusions and future work are presented.

2 Related Work

FSS techniques are typically classified into three groups: wrapper methods, filter methods and embedded methods [12]. The main difference between them is that wrapper

and embedded methods are specific for the used classifier, while filter methods are independent of the employed classifier. Our work focuses on filter methods because, given that they do not depend on any classifier to select the variables, they are more general, less prone to overfitting and computationally cheaper than wrapper and embedded methods [11]. Basically, the objective of the filter FSS methods is to find the minimal subset of original features that retains the information contained in the whole set of features [11] for the purpose of classification. However, as explained in [13], feature selection is not an easy task, due to the complicated interactions that could occur between features.

While the recent literature accounts for an important number of works in filter methods for the problem of multivariate time series forecasting, this is not the case for the classification problem.

All of the filter methods for the problem of multivariate time series forecasting try to find the subset of predictor TS that best improves the prediction accuracy of a target TS. In [14], different FSS techniques based on Pearson correlation, Spearman correlation, Granger causality and mutual information are analyzed. The presented methods select a subset of predictor TS by pairwise comparing them with the target TS. However, they do not consider the relationships that may exist between selected predictor TS, thus redundancies are not discarded. Dealing with possible redundancies, a representative FSS example is presented in Motrenko et al. [15]. It proposes a Quadratic Programming Feature Selection (QPFS) method that selects a subset of predictor features by solving a quadratic problem that minimizes correlation between features while maximizing feature relevance. The shortcoming of this approach is the computation of the close to singular similarity matrix and the computational cost of the solution of the problem. The existence of a variety of studies that use information theory to develop feature selection methods in TS prediction is also worth mentioning. For example, in Karevan et al. [16], a feature selection method that tries to select the subset with the minimum conditional sample entropy of the target variable is presented. Here, a clustering-based sample entropy that is calculated applying the Heaviside kernel is used to perform feature selection. Meanwhile, Liu et al. [17] uses a mutual information criterion as a filter method. In this case, in order to estimate the mutual information between TS, Kraskov et al.'s [18] method is followed. However, how the method is modified to deal with TS is not detailed. Recently, González-Vidal et al. [19] proposed a method that, before applying different feature selection methods, removes the temporal ordering of the series and generates a new set of predictor variables by concatenating vectors of measurements at different timesteps.

Leaving the forecasting problem aside, for the MTS classification task, specific FSS methods have also been developed [20, 21]. Most of these methods are based on transforming the original TS using a different representation. For instance, methods such as shapelets [22], symbolic dynamic methods [23], pseudo-observations [24] are implemented, while others extract features as graph-based features [25], pairwise mutual information [26] (no details are given for the computation of the MI between TS) or correlations [27]. Recently, in [28], for the extraction and the selection of relevant and non-redundant multivariate ordinal patterns for classification, a technique called *Ordex* is presented. In addition, Bondi et al. [29] developed a method that first obtains different representations of the time series using derivatives, cumulative sums, autocorrelation between values of the signal at different time-stamps, and power spectrum. Next, the new uninformative features are filtered by taking into account the balance between their complexity and their informativeness. In summary, all these methods transform the original MTS data by generating new features of which a subset is se-

lected for classification.

To carry out the transformations of the features, it is necessary to provide all the original features, hence they have to be available and cannot be discarded. Apart from that, an issue that may result from these transformations is the lost of part of the information that time series have, for example the temporal information or the information about the relationship between the original series.

Another drawback in some FSS methods proposed in MTS is the lack of use of class information [27] and the non-detection of redundancies between TS [30]. In [31], the previously mentioned problems are solved by selecting the most relevant features for classification and the least redundant ones. Indeed, this proposal is the only one in the literature, to the best of our knowledge, that selects a subset of the original univariate time series, without transforming the series into a different representation as in the previously mentioned approaches. The authors calculate Pearson's correlation to identify linear relationships and the Symmetrical uncertainty filter, which is based on entropy and information gain, for evaluating the non-linear relationships. However, symmetrical uncertainty is only suitable for discrete data, so they previously discretize the time series using an unspecified discretization method. This may result in a loss of useful information of the time series that may negatively affect the selection process and, consequently, the classification results. Additionally, the lack of details about the discretization process in the paper and the unavailability of the code prevents us from reproducing the results and comparing our method with it.

Therefore, to improve these approaches, a mutual information based TS selection method for MTS classification is proposed. In our approach, we follow the Brown et al. [10] framework, which focuses on classification and is based on Shannon Entropy [32].

3 Proposed Method

This section presents the proposed time series subset selection method based on mutual information, which will be used to solve the multivariate time series classification problem.

Let us define a multivariate time series as a matrix of $n \times d$, where n is the number of time steps when d different variables are measured. By itself, each measure across time steps is an univariate time series $ts^i = [a_1^i, \ldots, a_n^i]$, for i in $\{1, \ldots, d\}$. So, a multivariate time series will be denoted as:

$$MTS = \begin{bmatrix} ts^1 \\ ts^2 \\ \vdots \\ ts^d \end{bmatrix} = \begin{bmatrix} a_1^1 & a_2^1 & \cdots & a_n^1 \\ a_1^2 & a_2^2 & \cdots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^d & a_2^d & \cdots & a_n^d \end{bmatrix}$$

In a supervised classification problem, m instances of the problem are observed. For each instance, a class label is assigned. After training a classification model with the observed instances, new instances can be classified. Thus, the supervised classification problem that we have studied is presented in Figure 1.

In this paper, we propose a filter FSS method specifically designed for the problem of supervised MTS classification explained above. As can be seen in Figure 2, we will

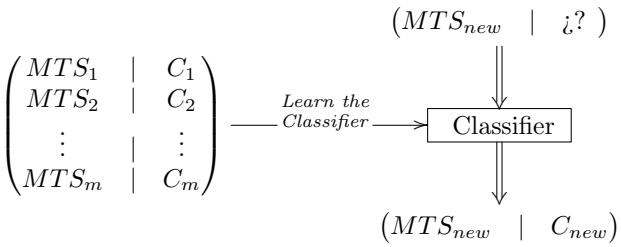

Figure 1: The studied MTS supervised classification problem.

depart from a labeled MTS dataset, and with our method, we will select a subset of the original univariate time series. In particular, our FSS method is an adaptation of the

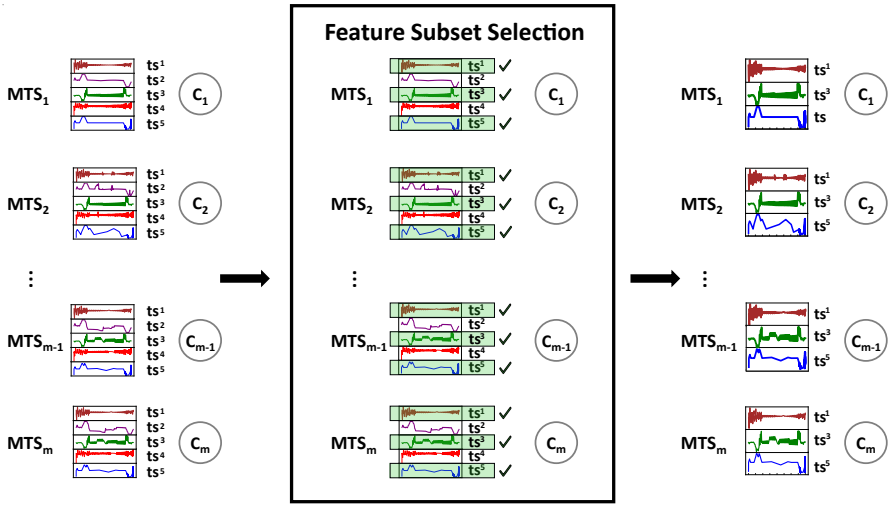

Figure 2: Diagram of the problem of feature subset selection that selects a subset of time series for a multivariate time series classification scenario.

probabilistic framework proposed by Brown et al. [10] for a time series. This method considers a score function based on MI to measure the relevance of each subset of features. In particular, for the case of TS, the score function can be written as follows:

$$\begin{aligned} J(S) = & \sum_{i \in S} I(TS^i; C) \\ & - \beta \sum_{j=1}^{z-1} \sum_{k=j+1}^{z} I(TS^j; TS^k) \\ & + \gamma \sum_{j=1}^{z-1} \sum_{k=j+1}^{z} I(TS^j; TS^k \mid C) \end{aligned} \tag{1}$$

where I is the mutual information, TS are time series, C is the class variable and S a

subset of TS, such that $S = \{TS^1, \ldots, TS^z\}$. The objective is to obtain the subset of TS that maximizes the score function J. Thus, the selected TS will be those that, at the same time, share high information with the classification variable (first MI term), share low information between them (second MI term) and those that are more related to each other given a class (third MI term).
To adjust the relevance given to each of these terms, the scoring function parameters β and $\gamma \in [0, 1]$ are set. In the literature, different criteria exist depending on the values assigned to β and γ. Based on the results of Brown et al. [10], the most common linear criteria are:

- Conditional Mutual Information (CMI): $\beta = 1$ and $\gamma = 1$. All the terms are equally relevant for the score function.
- Joint Mutual Information (JMI): $\beta = \frac{1}{|S|}$ and $\gamma = \frac{1}{|S|}$. The relevance of the mutual information between TS and the mutual information between TS conditioned to the classification variable decreases with $|S|$.
- Minimum-Redundancy Maximum-Relevancy (MRMR): $\beta = \frac{1}{|S|}$ and $\gamma = 0$. Class-conditional mutual information between TS is irrelevant, while the relevance of the mutual information between TS decreases proportionally with $|S|$.

The key point of the score function J (Equation (1)) is the computation of the mutual information terms involved. Contrary to the case of non-temporal vector-value features where many estimation methods have been proposed in the literature [9, 33, 34], the computation of the mutual information between two time series or between a time series and a class label C is an unsolved problem.
Two possible alternatives can be used to deal with this problem:

1. Assume a probabilistic model for the time series, such as, Autoregressive Moving Average Models (ARMA) or Gaussian Process models [35] and then, calculate information theory quantities departing from the models [36, 37].
2. Adapt to the time series case, those methods that do not assume any specific probability distribution model in the data, such as, [18, 38, 39, 40].

The first approach has some drawbacks, for example, the selected model may not always fit the data properly. Moreover, the calculations to obtain the MI are complex [36, 37]. Therefore, we have chosen the second approach. The methods for the computation of each mutual information term will be described in the next section.

Having defined the score function J (Equation (1)), algorithms for searching the subset of TS that maximizes it are now needed. Different searching algorithms could be used for selecting the most adequate time series subset. The most common in the FSS literature are Forward, Backward and Stepwise iterative algorithms [41].

4 Mutual Information Estimation Methods

In this section, MI estimation methods adaptations that allow the computation of the previously mentioned score function terms are described.

The score function J in Equation (1) requires the computation of the following MI terms: 1) $I(TS^i;TS^j)$, 2) $I(TS;C)$ and 3) $I(TS^i;TS^j \mid C)$. To compute them, methods that do not assume any probability distribution model in the data are considered [18, 38, 39, 40]. All of these methods estimate the information shared between vector-value variables using a k-nearest neighbor strategy. Our approach attempts to adapt these techniques to the time series scenario.

Because of the temporal information that the time series contain, a TS specific measure has been applied in the computation of the k-nearest neighbor. Specifically, Dynamic Time Warping (DTW) dissimilarity is proposed. DTW is designed for TS and it allows the alignment of TS in a non-linear way by minimizing the distance between them [3]. Thanks to DTW, MI terms will account for the temporal information contained in the time series. The adapted methods for computing the three MI terms are explained in the sections that follow. To contribute to understanding, we present them in a different order than in the score function J (Equation (1)).

4.1 Mutual information between a time series $I(TS^i;TS^j)$

To estimate the term $I(TS^i;TS^j)$, Kraskov et al.'s [18] method is followed. This method estimates the MI between two continuous random variables using the k-nearest neighbor to quantify the information shared. To apply this method to TS, it is adapted as follows. First, we assume that TS^i and TS^j are random variables whose realizations are time series. We assume we have an m-size sample, $\{(ts_1^i, ts_1^j), (ts_2^i, ts_2^j), \ldots, (ts_m^i, ts_m^j)\}$ where $ts_p^i \in TS^i$ and $ts_p^j \in TS^j$ for $p \in \{1, \ldots, m\}$. Note that any element of the sample ts_p^l is a time series for $l \in \{i, j\}$. For each time series ts_p^l of the sample, for $l \in \{i, j\}$, let the value $\xi_{ts_p^l}$ be the distance from ts_p^l to its k^{th} nearest neighbor TS in the sample. As commented previously, we use DTW to calculate this value. Then, using $\xi_{ts_p^i}$ and $\xi_{ts_p^j}$, $\xi(p)$ value is defined as $\xi(p) = \max(\xi_{ts_p^i}, \xi_{ts_p^j})$. From this, the value

$$\nu_{ts_p^l} = \left| \left\{ q \mid q \in \{1, \ldots, m\} - \{p\},\ DTW(ts_q^l, ts_p^l) < \xi(p) \right\} \right|$$

is determined for $l \in \{i, j\}$. Again, the DTW is used as a distance between TS. After these definitions, the mutual information between two time series TS^i and TS^j is estimated by the adaptation of the method proposed in Kraskov et al. [18], as follows:

$$\begin{aligned} I(TS^i;TS^j) = \ & \Psi(k) + \Psi(m) \\ & - \frac{1}{k} - \left(\frac{1}{m} \sum_{p=1}^{m} \left(\Psi(\nu_{ts_p^i}) + \Psi(\nu_{ts_p^j}) \right) \right) \end{aligned} \tag{2}$$

where Ψ is the digamma function, which can be calculated recurrently as $\Psi(x+1) = \Psi(x) + 1/x$ with $\Psi(1) = -C$, where $C = 0.5772156...$ is the Euler-Mascheroni constant. TS^i and TS^j will share more information when, for all of the instances, the k nearest time series samples for both random variables keep the same distance proportion.

4.2 Mutual information between time series and the classification variable $I(TS;C)$

To estimate the first term of the score function, $I(TS;C)$, three different methods of MI estimation between a continuous and a discrete variable are considered. They

will be adapted to measure the information shared between a TS and the associated classification variable. The methods that will be followed are the Ross method [38], Bulinski and Kozhevin method [39] and Coelho et al. method [40].

The Ross method [38] was designed to estimates the MI between a discrete random variable and a continuous random variable. It uses a similar approach to the one used in Kraskov et al. [18]. The proposed adaptation to deal with TS is as follows: let TS be defined as in previous section and let C be a discrete classification variable taking values in a finite set. A sample of size m, $\{(ts_1, c_1), (ts_2, c_2), \ldots, (ts_m, c_m)\}$, is given to estimate MI. From the sample, ν_{c_p} value is defined as the number of sample pairs whose class value is equal to c_p. Then, within this subset of sample pairs, whose class value is equal to c_p, the distance d_p from time series sample ts_p to its k^{th} nearest neighbor is determined. DTW is used as a distance between TS. Then, ν_{ts_p} value is set to be, $\nu_{ts_p} = |\{\ q \mid q \in \{1, \ldots, m\} - \{p\},\ DTW(ts_q, ts_p) \leq d_p\}|$. After these definitions, the adaptation of the Ross method [38] to estimate mutual information between a discrete random variable and a TS is as follows:

$$I(TS; C) = \Psi(k) + \Psi(m) - \frac{1}{m}\left(\sum_{p=1}^{m}\left(\Psi(\nu_{c_p}) + \Psi(\nu_{ts_p})\right)\right) \tag{3}$$

where Ψ is again the digamma function. On this occasion, the closer the distance between series of the same class, the higher the mutual information.

Similarly to the previous case, in Bulinski and Kozhevin [39] propose a method to estimate the conditional entropy for a discrete random variable given a continuous random variable. Following this approach, the proposed adaptation allows us to estimate the conditional entropy, $H(C|TS)$, of the classification variable C given a TS. Then, $I(TS; C)$ is calculated using the relationship that exists between MI and entropy, $I(TS; C) = H(C) - H(C \mid TS)$.

Let TS and C be as previously defined. A sample of size m, $\{(ts_1, c_1), (ts_2, c_2), \ldots, (ts_m, c_m)\}$, is given. Using this sample, the conditional entropy $H(C|TS)$ is estimated. First, for each time series ts_p of the sample, ξ_{ts_p} value that is the distance from ts_p to its k^{th} nearest neighbor TS in the sample is set. DTW dissimilarity is applied. Then, using ξ_{ts_p}, ν_p value is defined as $\nu_p = \left|\{\ q \mid q \in \{1, ..., m\} - \{p\},\ c_q = c_p, DTW(ts_q, ts_p) \leq \xi_{ts_p}\}\right|$. Then, $H(C|TS)$ is estimated adapting the method developed by Bulinski and Kozhevin [39] as:

$$H(C \mid TS) = \log(k) - \frac{1}{m}\sum_{p=1}^{m}\log\left(\nu_p + 1\right) \tag{4}$$

Once $H(C \mid TS)$ is estimated, only by estimating $H(C)$ for the discrete variable, $I(TS; C)$ is obtained.

In the case of Coelho et al.'s [40] method, they use the Kozachenko-Leonenko entropy estimator [18] to estimate the entropy of a continuous random variable and the conditioned entropy given a classification variable. Considering this approach, an adaptation is presented to estimate both entropy terms, $H(TS)$ and $H(TS|C)$, for a time series. Then, $I(TS; C)$ is calculated with the equivalent relationship between MI and entropy to the one presented in the previous method but conditioned to C. It is supposed to have a sample of size m, $\{(ts_1, c_1), \ldots, (ts_m, c_m)\}$. For each time series

ts_p, ξ_{ts_p} value is the distance from ts_p to its k^{th} nearest neighbor TS in the sample is set. To calculate distances between TS, DTW is applied. Following Coelho et al.'s [40] method, the adaptation for calculating entropy of time series is presented as:

$$H(TS) = -\Psi(k) + \Psi(m) + \frac{1}{m}\left(\sum_{p=1}^{m} \log(2 \cdot \xi_{ts_p})\right) \tag{5}$$

Then, $H(TS|C)$ is defined as:

$$H(TS|C) = \sum_{s=1}^{k} H(TS \mid c_s) \cdot p(c_s) \tag{6}$$

where $p(c)$ is the probability distribution function of C that takes values in $\{c_1, \ldots, c_k\}$. $H(TS)$ and $H(TS|c_s)$ are computed with the same method but the second term uses a subsample of the given sample, restricting it to the instances with class value equal to c_s. After estimating both terms $H(TS)$ and $H(TS \mid C)$, $I(TS;C)$ is obtained.

4.3 Conditional mutual information between time series $I(TS^i; TS^j \mid C)$

Finally, to estimate the third term in the score function, $I(TS^i; TS^j \mid C)$, we base on the following equation:

$$I(TS^i; TS^j \mid C) = \sum_{s=1}^{k} I(TS^i; TS^j \mid c_s) \cdot p(c_s) \tag{7}$$

where $I(TS^i; TS^j \mid c_s)$ is estimated with the previously described adapted method for computing $I(TS^i; TS^j)$ in Equation (2), but restricted to those instances in which the classification variable C is equal to the class label c_s [12]. The probability of each class must also be estimated before we can estimate MI.

5 Experimental Framework

This section will present all of the requirements for carrying out the evaluation of the proposed method in MTS classification problems.

To evaluate our time series subset selection method in the solutions of MTS classification problems, some of the benchmark datasets provided in [42] are used. Large datasets requiring large computational resources and datasets with less than 6 dimensions for which FSS methods become meaningless are discarded. The properties of the selected datasets are summarized in Table 1.

Before applying the proposed time series subset selection method to the datasets, some components and parameters need to be set. In particular, we consider different MI estimation methods, different score functions, different parameters for MI estimation and different search algorithms. The following algorithms and parameters are used to define an instance of the proposed framework for time series subset selection:

- **Methods for MI estimation between a TS and the classification variable:** the Ross method (1) [38], Bulinski and Kozhevin method (2) [39] and Coelho et al. method (3) [40] are considered.

Table 1: Properties of selected datasets

Dataset	Dimensions (d)	TS length (n)	Classes (c)
ArticularyWordRecognition	9	144	25
BasicMotions	6	100	4
Cricket	6	1197	12
DuckDuckGeese	1345	270	5
EigenWorms	6	17984	5
FingerMovements	28	50	2
HandMovementDirection	10	400	4
Heartbeat	61	405	2
JapaneseVowels	12	29	9
LSST	6	36	14
MotorImagery	64	3000	2
NATOPS	24	51	6
PEMS-SF	963	144	7
Phoneme	11	217	39
RacketSports	6	30	4
SelfRegulationSCP1	6	896	2
SelfRegulationSCP2	7	1152	2

- **Score function criteria:** CMI (C), JMI (J) and MRMR (M).
- **Distances used in the methods for MI estimation:** Euclidean distance (EU), which ignores the temporal information of the time series, and dynamic time warping dissimilarity (DTW), which includes the temporal information, are considered.
- ***k* for k-nearest neighbor in MI estimation methods:** values 1, 3, 6, 10, 13, 16, 20 for k are considered.
- **Searching algorithms:** forward (F), backward (B) and stepwise (S).

Each subset returned by the proposed FSS method will be evaluated by means of a 1-NN classifier with DTW dissimilarity. Despite its simplicity, it achieves competitive results and in the literature it is considered as a benchmark [4, 43]. In the case of MTS, two generalizations of the DTW are commonly used: the dimension independent DTW (DTW_I) and the dimension dependent DTW (DTW_D) [43, 44]. DTW_I is calculated as the sum of the DTW distances in each dimension. In contrast, for DTW_D, the values of all the time series at time t are considered as a vector. Then, at each pair of time steps, the Euclidean distance is computed between the corresponding vectors and the DTW_D is calculated departed from these values. Both classification alternatives, 1-NN with DTW_D and 1-NN with DTW_I, will be considered in the following experiments.

All of the MTS classification problems that we have used come divided into two datasets, one for training and the other for testing. The training dataset will be used to apply our FSS method. The testing dataset will then be filtered with the output of the FSS method. Finally, the selected time series will be used to evaluate the classification and also the FSS method. Classification accuracy has been used to evaluate the performance of a classifier. In particular, the classification accuracy obtained will be compared with the accuracy obtained when the classification is performed with all available TSs. Therefore, we will examine whether our method returns TS subsets that improve classification accuracy.

6 Results and Discussion

In this section, the obtained results after running the experiments are discussed.

For each dataset, all of the possible parameters configurations are examined. In total, for each dataset, 756 experiments have been run (3 methods for MI estimation between TS and class variable x 3 score functions x 2 MI estimation distances x 7 parameters k in MI estimation x 3 searching algorithms x 2 classification alternatives).

Once the global best results are analyzed, we evaluate the sensitivity of the method to different characteristics and parameters of the algorithm. Specifically, we analyze the results considering the different options for the score function criterion, distance used for MI estimation, the parameter k used for MI estimation, the search algorithm and the classification alternatives. To this end, we calculate the percentage of features selected by each configuration of the method with respect to the initial number of features in the dataset (% of selected TS). We also measure the improvement in accuracy obtained when the classification is performed only with the selected features as opposed to the classification with all the initial features (% Improvement Acc).

The following six sections will present these results.

6.1 Global best results

The objective of our first analysis is to try to find the parameter combinations of the proposed method that work the best for all of the datasets.

For each dataset, we consider the best accuracy obtained by all the different parameter combinations. In case of ties, we identify those parameter combinations that produce the subset with the lowest number of TS as being the best.

The best results for each dataset along with the summary of the parameter configurations that return these results are presented in Table 2. The displayed columns are, from left to right: 1) the dataset, 2) the best accuracy obtained after applying the time series subset selection methods versus the accuracy obtained with all the available TS, 3) the number of TS that the best selected subset has versus the number of possible time series in the dataset, 4) the number of parameter combinations that obtain the best accuracy versus the number of experiments performed, 5) the classifier applied (1-NN with dependent DTW or 1-NN with independent DTW), and from 6) to 10) columns the parameters and alternatives of the proposed method where the best value was reached, in the same order as defined in the previous section (Section 5).

Several conclusions can be extracted from Table 2. First, although the percentage of improvement depends on the dataset, in almost all of the cases, the selected subset of time series strongly improves the accuracy reached when all the series are used (except in ArticularyWordRecognition with the 1-NN with DTW_D and, JapaneseVowels, LSST and RacketSports with the 1-NN with DTW_I).

Second, in addition to improving accuracy, our method is able to dramatically reduce the number of series with regards to the original subset. This demonstrates that the proposed method has a remarkably effective performance.

However, the percentage of parameter configurations that yield the best results are low and there is apparently no clear parameter combination pattern that works better than the rest of them for all of the datasets. Consequently, selecting a successful parameter configuration is not an easy task.

With regard to the methods for MI estimation between a TS and the classification

Table 2: Best results for each dataset. The table contains those combinations of parameters that achieve the highest accuracy for each dataset together with the accuracy, the number of selected TS and the number of optimal configurations. In case of accuracy ties, the combinations that select the smallest subset of TS are established as the best results.

Dataset	Best accuracy	Nº of TS s.	Opt. config.	Classif. altern.	MI class 1 / 2 / 3	Score crit. C / J / M	MI dist. EU / DTW	k-NN MI class 1/3/6/10/13/16/20	Search alg. F / B / S
Articulary	99 / 99	7 / 9	9 / 378	DTW_D	0 / 9 / 0	0 / 5 / 4	3 / 6	0 / 0 / 2 / 4 / 3 / 0 / 0	1 / 4 / 4
WordRec.	98.7 / 98.3	8 / 9	9 / 378	DTW_I	3 / 6 / 0	0 / 1 / 8	3 / 6	0 / 0 / 1 / 2 / 6 / 0 / 0	5 / 2 / 2
Basic	100 / 87.5	1 / 6	11 / 378	DTW_D	3 / 8 / 0	4 / 3 / 4	11 / 0	3 / 1 / 2 / 3 / 1 / 1 / 0	11 / 0 / 0
Motions	100 / 100	1 / 6	11 / 378	DTW_I	3 / 8 / 0	4 / 3 / 4	11 / 0	3 / 1 / 2 / 3 / 1 / 1 / 0	11 / 0 / 0
Cricket	100 / 100	3 / 6	1 / 378	DTW_D	0 / 1 / 0	1 / 0 / 0	0 / 1	0 / 0 / 0 / 0 / 1 / 0 / 0	0 / 0 / 1
	100 / 98.61	3 / 6	1 / 378	DTW_I	0 / 1 / 0	1 / 0 / 0	0 / 1	0 / 0 / 0 / 0 / 1 / 0 / 0	0 / 1 / 0
DuckDuck	32.5 / 25	201/ 1345	1 / 378	DTW_D	0 / 1 / 0	1 / 0 / 0	1 / 0	0 / 1 / 0 / 0 / 0 / 0 / 0	0 / 0 / 1
Geese	35 / 17.5	201/ 1345	4 / 378	DTW_I	4 / 0 / 0	0 / 4 / 0	4 / 0	0 / 0 / 0 / 1 / 1 / 1 / 1	0 / 0 / 4
Eigen	74 / 65.7	4 / 6	33 / 378	DTW_D	15/ 12/ 6	27 / 3 / 3	27 / 6	9 / 4 / 4 / 10 / 3 / 3 / 0	10 / 13 / 10
Worms	67.2 / 62.6	5 / 6	41 / 378	DTW_I	27/ 14/ 0	0 / 18 / 23	0 / 41	3 / 3 / 8 / 9 / 6 / 6 / 6	14 / 13 / 14
Finger	67.0 / 51.0	1 / 28	15 / 378	DTW_D	12 / 3 / 0	5 / 5 / 5	12 / 0	0 / 0 / 0 / 3 / 3 / 3 / 6	15 / 0 / 0
Mov.	67.0 / 51.0	1 / 28	15 / 378	DTW_I	12 / 3 / 0	5 / 5 / 5	12 / 0	0 / 0 / 0 / 3 / 3 / 3 / 6	15 / 0 / 0
HandMov.	33.3 / 19.7	2 / 10	34 / 378	DTW_D	10/ 24/ 0	12 / 9 / 13	20 / 14	0 / 5 / 8 / 6 / 3 / 6 / 6	0 / 0 / 34
Direction	34.0 / 30.6	2 / 10	1 / 378	DTW_I	0 / 0 / 1	0 / 0 / 1	1 / 0	0 / 1 / 0 / 0 / 0 / 0 / 0	0 / 0 / 1
Heartbeat	72.7 / 66.8	5 / 61	1 / 378	DTW_D	1 / 0 / 0	0 / 0 / 1	0 / 1	0 / 0 / 0 / 0 / 0 / 0 / 1	1 / 0 / 0
	70.7 / 69.8	3 / 61	1 / 378	DTW_I	1 / 0 / 0	0 / 0 / 1	1 / 0	0 / 0 / 1 / 0 / 0 / 0 / 0	1 / 0 / 0
Japanese	81.9 / 80.5	10 / 12	9 / 378	DTW_D	9 / 0 / 0	0 / 9 / 0	3 / 6	3 / 3 / 3 / 0 / 0 / 0 / 0	3 / 3 / 3
Vowels	81 / 81	10 / 12	93 / 378	DTW_I	6/ 3/ 84	0 / 51 / 42	48 / 45	15/ 12/ 12/ 15/ 15/ 12/ 12	31 / 31 / 31
LSST	48.1 / 44.2	2 / 6	6 / 378	DTW_D	4 / 0 / 2	4 / 0 / 2	4 / 2	5 / 1 / 0 / 0 / 0 / 0 / 0	0 / 0 / 6
	51.2 / 51.2	6 / 6	105/ 378	DTW_I	9/ 12/ 84	0 / 63 / 42	48 / 57	21/ 21/ 15/ 12/ 12/ 12/ 12	35 / 35 / 35
Motor	63 / 44	2 / 64	1 / 378	DTW_D	0 / 1 / 0	0 / 0 / 1	0 / 1	0 / 0 / 0 / 1 / 0 / 0 / 0	0 / 0 / 1
Imagery	59 / 58	2 / 64	3 / 378	DTW_I	0 / 3 / 0	0 / 0 / 3	2 / 1	0 / 0 / 0 / 1 / 1 / 1 / 0	0 / 0 / 3
NATOPS	81.7 / 77.2	3 / 24	7 / 378	DTW_D	0 / 0 / 7	7 / 0 / 0	0 / 7	0 / 0 / 0 / 2 / 2 / 2 / 1	4 / 0 / 3
	79.4 / 74.4	11 / 24	1 / 378	DTW_I	1 / 0 / 0	0 / 0 / 1	1 / 0	1 / 0 / 0 / 0 / 0 / 0 / 0	0 / 0 / 1
PEMS-SF	95.4 / 81.5	33 / 963	1 / 378	DTW_D	0 / 0 / 1	1 / 0 / 0	1 / 0	1 / 0 / 0 / 0 / 0 / 0 / 0	0 / 1 / 0
	96.5 / 82	33 / 963	1 / 378	DTW_I	0 / 0 / 1	1 / 0 / 0	1 / 0	1 / 0 / 0 / 0 / 0 / 0 / 0	0 / 1 / 0
Phoneme	15.6 / 15	6 / 11	6 / 378	DTW_D	3 / 3 / 0	6 / 0 / 0	0 / 6	0 / 0 / 0 / 2 / 2 / 1 / 1	2 / 3 / 1
	15.6 / 15	6 / 11	6 / 378	DTW_I	3 / 3 / 0	6 / 0 / 0	0 / 6	0 / 0 / 0 / 2 / 2 / 1 / 1	2 / 3 / 1
Racket	88.8 / 88.2	4 / 6	2 / 378	DTW_D	2 / 0 / 0	0 / 2 / 0	2 / 0	0 / 0 / 2 / 0 / 0 / 0 / 0	1 / 0 / 1
Sports	87.5 / 87.5	6 / 6	157/ 378	DTW_I	49/ 24/ 84	0 / 92 / 65	55 / 102	29/ 27/ 22/ 21/ 19/ 19/ 20	50 / 55 / 52
SelfReg.	81.6 / 77.5	2 / 6	18 / 378	DTW_D	5 / 8 / 5	10 / 4 / 4	8 / 10	3 / 2 / 0 / 2 / 3 / 2 / 6	1 / 17 / 0
SCP1	80.9 / 79.2	2 / 6	11 / 378	DTW_I	3 / 6 / 2	8 / 3 / 0	3 / 8	0 / 0 / 7 / 4 / 0 / 0 / 0	0 / 11 / 0
SelfReg.	56.1 / 51.1	2 / 7	27 / 378	DTW_D	6 / 9 / 12	11 / 5 / 11	18 / 9	2 / 2 / 6 / 0 / 7 / 3 / 7	0 / 0 / 27
SCP2	52.2 / 51.7	2 / 7	4 / 378	DTW_I	0 / 2 / 2	2 / 0 / 2	0 / 4	2 / 0 / 0 / 0 / 0 / 2 / 0	0 / 0 / 4

variable, it is observed that there is no method that clearly improves the results of the others. Depending on the dataset, the best results are obtained with different methods, but without following any obvious pattern.

6.2 Evaluation of the score function criterion

In the experiments, the CMI, JMI and MRMR score functions are tested. The intention is to extract information about which is the most appropriate for TS subset selection.

For each dataset and criterion, the average of the "% Improvement Acc" and the average of the "% of selected TS" is calculated. Then, in Figure 3, for each criterion, a scatter plot between both measures is presented (the further to the top-left, the better). The average "% of selected TS" for each criterion will depend on the output of all the experiments using that criterion in the FSS method. Each dataset is represented using a different shape. The two different colors represent the distances used for MI estimation, DTW dissimilarity and Euclidean distance. Because they influence the final score function together with the criterion, we add this information to the dispersion graph.

Figure 3 shows that disparities exist in the obtained results. On the one hand, it can be determined that among the three methods, the CMI, on average, selects the smallest relative sets of TS. That is, the points appear more to the left than for the rest of the criteria. However, in terms of the average "% Improvement Acc", the CMI criterion, for certain datasets such as Japanese Vowels, Heartbeat, Articulary Word Recognition, PEMS-SF and LSST, is worse than the others. A possible reason is that the CMI criterion is more restrictive with the features that it selects and, consequently, some classification information is lost, resulting in no improvement in the accuracy.

On the other hand, taking into account the average "% of selected TS", the JMI criterion returns larger subsets than the others. If we analyze the score function J when the JMI criterion is used (see Section 3), we conclude that a possible reason could be that the weight of the second and third terms of J diminish their relevance in the score function when the selected TS subset S grows.

Finally, with the exception of Japanese Vowels, LSST and Racket Sports datasets, MRMR is the criterion that obtains the best average "% Improvement Acc". However, on average, it selects a higher percentage of TS than the CMI.

6.3 DTW vs Euclidean distance in mutual information estimation

The main adaptation proposed in MI estimation methods for time series is the use of the DTW dissimilarity to consider the temporal information of the TS. To validate the performance of our adaptation, the results obtained using the DTW dissimilarity to estimate the MI terms are compared with those obtained using the Euclidean distance.

The experiments that achieve an improvement in accuracy after applying our FSS method when compared to using all the time series are first filtered. Then, for each dataset, we count the number of these configurations that use DTW dissimilarity and those that use Euclidean distance when estimating the MI. Finally, the obtained results are segmented by the selected classification alternative (DTW_D or DTW_I). This information is shown in Table 3, where the distance measure that obtains the highest number of configurations for each dataset and classification alternative is highlighted.

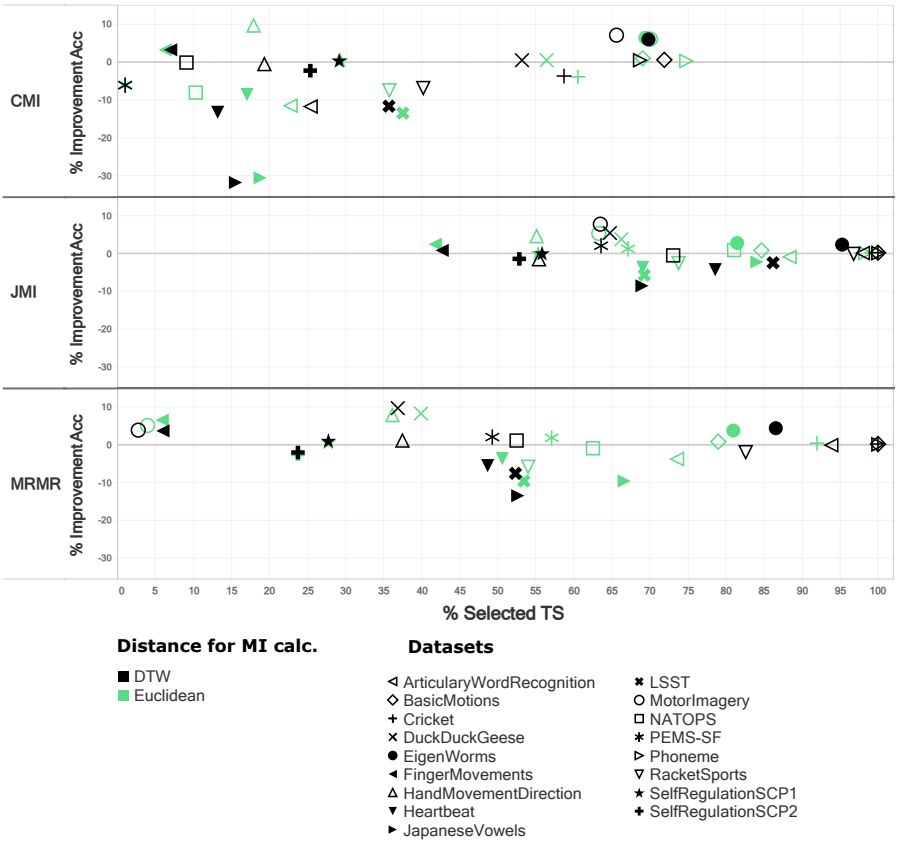

Figure 3: Dispersion table with "% of selected TS" and "% Improvement Acc" for each selection criterion and distance used for MI estimation in the different datasets. Note that the further to the top-left, the better.

In Table 3 this crossed information is presented for each dataset. The higher number of configurations between the Euclidean distance and the DTW dissimilarity in MI estimation are highlighted for each dataset and classification alternative.

Table 3 provides more insight into how the applied distances affect the results. It can be seen that for the majority, the number of parameter configurations that improve the accuracy using the DTW is higher than using the Euclidean distance. This result validates the main adaptation proposed to add the temporal information of the series in the computation of the MI.

6.4 Sensitivity of the k parameter in the mutual information estimation

The selection of the k parameter in the estimation of the MI quantities is another aspect to consider in the application of this time series selection method. It can vary

Table 3: Number of parameter configurations that improve the accuracy for each dataset divided by the classification alternatives and the distances used for MI estimation. The maximum number of combinations between MI computation distances and between classification alternatives are highlighted for each dataset.

Datasets	Classif. altern.	MI dist. EU	MI dist. DTW	Total
Artic.	DTW_D	94	**111**	**205**
Word	DTW_I	94	**105**	199
Recog.	**Total**	188	**216**	404
Basic	DTW_D	171	**178**	**349**
Motions	DTW_I	160	**179**	339
	Total	331	**357**	688
	DTW_D	149	**150**	299
Cricket	DTW_I	149	**154**	**303**
	Total	298	**304**	602
Duck	DTW_D	80	**87**	167
Duck	DTW_I	183	**186**	**369**
Geese	**Total**	263	**273**	536
Eigen	DTW_D	182	**189**	**371**
Worms	DTW_I	165	**189**	354
	Total	347	**378**	725
Finger	DTW_D	121	**128**	249
Mov.	DTW_I	124	**139**	**263**
	Total	245	**267**	512
Hand	DTW_D	187	**188**	**375**
Mov.	DTW_I	**43**	42	85
Direction	**Total**	230	230	460
Heart	DTW_D	**101**	57	**158**
beat	DTW_I	**74**	43	117
	Total	**175**	100	275
Japanese	DTW_D	**60**	54	**114**
Vowels	DTW_I	**48**	45	93
	Total	**108**	99	207

Datasets	Classif. altern.	MI dist. EU	MI dist. DTW	Total
	DTW_D	81	**98**	**179**
LSST	DTW_I	48	**57**	105
	Total	129	**155**	284
Motor	DTW_D	189	189	**378**
Imagery	DTW_I	**4**	3	7
	Total	**193**	192	385
NATOPS	DTW_D	81	**128**	209
	DTW_I	125	**150**	**275**
	Total	206	**278**	484
PEMS-SF	DTW_D	105	**107**	212
	DTW_I	101	**113**	**214**
	Total	206	**220**	426
Phoneme	DTW_D	189	189	378
	DTW_I	189	189	378
	Total	378	378	756
Racket	DTW_D	57	**102**	**159**
Sports	DTW_I	55	**102**	157
	Total	112	**204**	316
SelfReg.	DTW_D	**183**	181	**364**
SCP1	DTW_I	85	**96**	181
	Total	268	**277**	545
SelfReg.	DTW_D	74	**83**	**157**
SCP2	DTW_I	26	**30**	56
	Total	100	**113**	213
Total		3777	**4041**	7818

between different values and finding the best could be time consuming. Consequently, different k values are tested to find how they affect the TS selection process.

In Figure 4, a dispersion graph is presented, where the average "% Improvement Acc" and average "% of selected TS" are crossed. For each dataset, a color and a shape are set. Depending on the k chosen for the MI estimation, the shape has a different size; that is, when $k = 1$ the shape will be smaller that when $k = 20$.

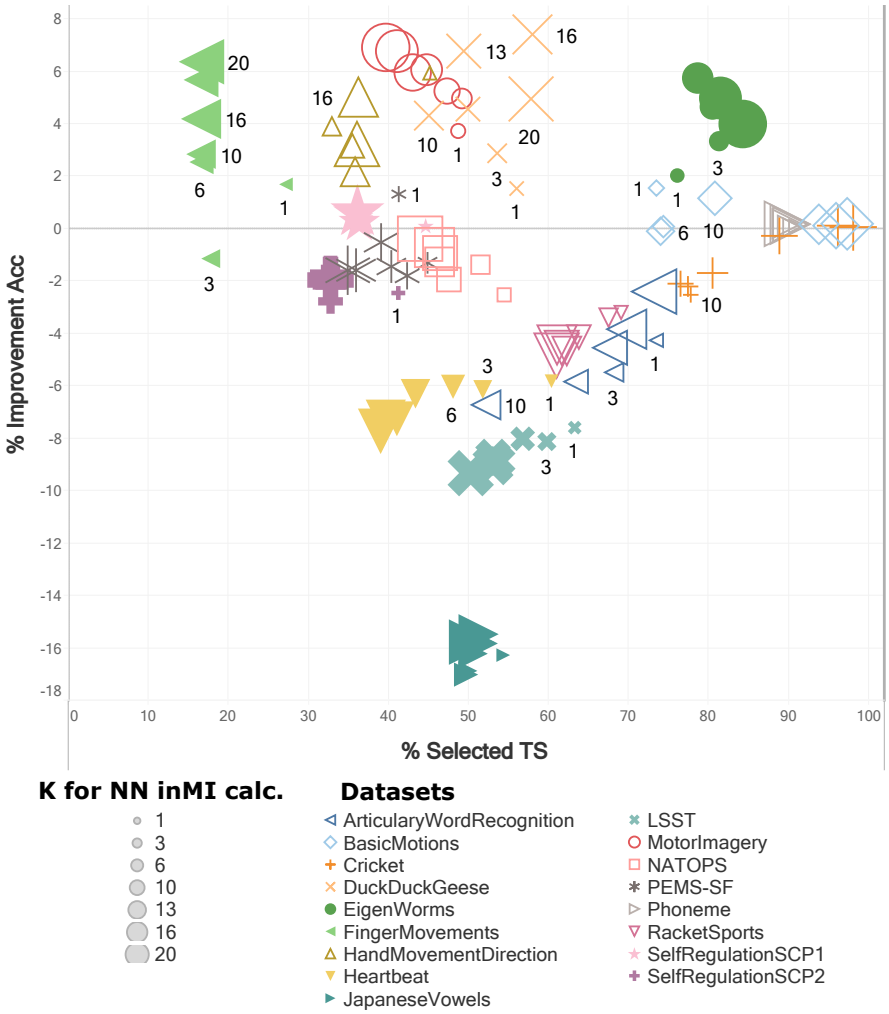

Figure 4: Dispersion graph with "% of selected TS" and "% Improvement Acc" for each k-value used in the nearest neighbor algorithm for the estimation of the mutual information in the different datasets. Each k-value has a different size in the shapes.

Analyzing Figure 4, it is observed that, except for ArticularyWordRecognition,

DuckDuckGeese and Finger Movements the choice of k has minimal influence on the "% Improvement Acc" because in that axis the points of the same dataset are grouped together. Hence, despite the fact that it has a higher influence for the "% of selected TS", it can be determined that the choice of parameter k, in general, has low influence in the performance of our FSS method and in the obtained results.

6.5 Evaluation of the searching algorithms

Due to the relevance that the searching algorithm could have in the process of time series subset selection, the objective is to analyze the performance of the applied searching algorithms. We will study the relationship that they have with the amount of selected time series and the reached classification accuracy.

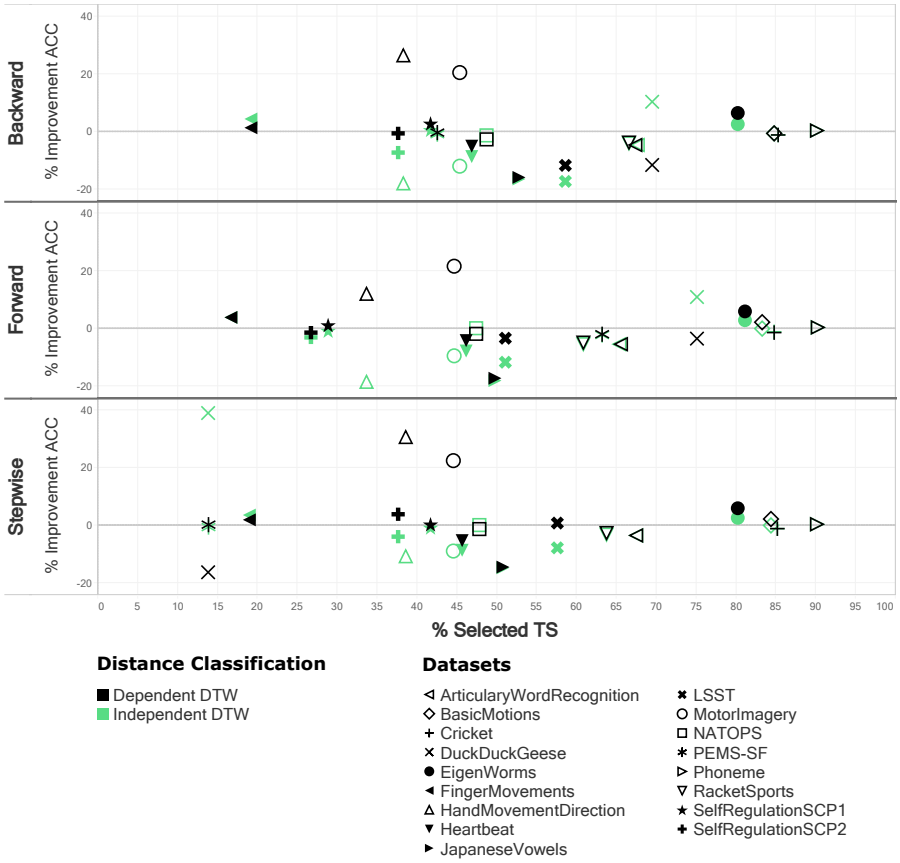

Figure 5: Dispersion table with "% of selected TS" and "% Improvement Acc" for each optimal searching algorithm used for selecting TS and each classification alternative in the datasets.

In Figure 5, a scatter graph is displayed, where the average "% Improvement

Table 4: Average "% Improvement Acc" of each dataset separated by the two classification alternatives (1-NN with DTW_D and 1-NN with DTW_I).

Datasets	Classification alternatives $\mathbf{DTW_D}$	$\mathbf{DTW_I}$
ArticularyWordRecognition	-4,74	-4,73
BasicMotions	1,25	-0,38
Cricket	-1,31	-1,20
DuckDuckGeese	-10,63	19,83
EigenWorms	5,95	2,48
FingerMovements	2,32	3,97
HandMovementDirection	22,83	-15,95
Heartbeat	-4,94	-8,38
JapaneseVowels	-15,97	-16,40
LSST	-4,77	-12,24
MotorImagery	21,41	-10,15
NATOPS	-2,05	-0,59
PEMS-SF	-0,69	-1,33
Phoneme	0,09	0,09
RacketSports	-3,98	-4,40
SelfRegulationSCP1	1,33	-0,41
SelfRegulationSCP2	0,56	-4,79

Acc" and average "% of selected TS" are crossed for each searching algorithm. Each dataset is represented using a different shape. The two different colors represent the classification alternatives applied, 1-NN with dependent DTW and 1-NN with independent DTW.

It can be observed in the figure that there are not substantial differences between the search algorithms in terms of accuracy and number of selected series.

6.6 $\mathbf{DTW_D}$ vs $\mathbf{DTW_I}$ in 1-NN classifier

While our FSS method is independent of the classifier used, we would like to analyze if there are any differences in the use of the DTW_D and the DTW_I in the 1-NN.

To examine the relationship between the used classification alternatives and the proposed TS selection method, the "% Improvement Acc" is calculated for each experiment. The average result is then computed for each classification alternative and each dataset. In Table 4, these results are presented.

It can be observed that the average "% Improvement Acc" derived from all the experiments is higher for the classification with DTW_D than for the DTW_I in almost all the evaluated datasets, except for Cricket, DuckDuckGeese, Finger Movements and NATOPS datasets. A possible explanation for this is that the 1-NN with dependent DTW considers all of the time series at once when calculating distances for classification and, therefore, it takes more advantage of removing those time series that are redundant. In any case, the differences in the average "% Improvement Acc" between both classifiers are not enough to yield any general conclusion, and we thus conclude that the average improvement in accuracy depends on the dataset.

7 Conclusion

The results of this paper show that the proposed TS subset selection method based on MI strongly increases the classification accuracy, while reducing the number of time series chosen to solve a MTS classification problem. Consequently, thanks to the realized experiments in multivariate time series classification, it has been demonstrated that our feature subset selection pre-process is a relevant step to improve the accuracy in MTS classification problems.

Several parameter combinations have been considered in the development of this time series subset selection method based on MI. The results of the method will be affected by the choice of parameter, distance and algorithm. In fact, no clear parameter pattern has been found that works correctly for all datasets. Nevertheless, some useful conclusions have been found for the selection of the best parameters. First, regarding to the distance used for the mutual information estimation, using the DTW dissimilarity instead of Euclidean distance leads to high probability of improving accuracy compared to classification using all features. This result validates the main adaptation proposed to add the temporal information of the series in the computation of the MI.

In addition, our results reveal that, after applying the proposed method, dependent DTW dissimilarity based 1-NN classification obtains a slightly higher percentage of improvement in accuracy when classifying the available TS than the independent DTW, in most cases. Hence, we can deduce that when using the DTW in the MI estimation and the DTW_D for the classification, generally, the probability of overcoming the accuracy obtained using all of the available time series is higher.

Moreover, considering all the experimental results, in most cases, when the CMI criterion is applied, our FSS method returns more reduced subsets of TS and, when the MRMR criterion is selected, the method obtains results with higher classification accuracy values. In conclusion, the choice of the selection criterion seems to affect our method in finding reduced subsets of TS that improve the accuracy with respect to the original set of TS. Even so, taking into account that the three criterion obtain good results (see the best results in Table 2), the choice of one or the other criterion is not entirely dramatic. Finally, it is inferred that the choice of the k that is used for MI estimation, as well as the searching algorithm implemented to find the optimal TS subset, are not as influential as the MI estimation techniques or the distance used.

Despite the work that is done and the conclusions that are drawn, there is still room for improvement.
Our future investigation will be focused on new non-linear score functions as an alternative to the proposed function. In addition, new ways to add temporal information of the time series for MI estimation methods will also be examined in the future. Finally, the next step in our research will be to modify the proposed time series subset selection method based on the mutual information for multivariate time series classification problem, which will allow it to cope with massive time series datasets.

Acknowledgments

The authors wish to express their thanks to the Basque Government for their financial support of this research through the Elkartek program under the DIGITAL project (Grant agreement No. KK-2019/00095) and under the 3KIA project. Any opinions, findings and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of funding agencies.

Jose A. Lozano is partially supported by the Basque Government through the BERC 2018-2021 program and IT1244-19 and by the Spanish Ministry of Science, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718, TIN2016-78365-R and PID2019-104966GB-I00.

References

[1] O. L. Hasna, R. Potolea, Time series - A taxonomy based survey, in: 13th IEEE International Conference on Intelligent Computer Communication and Processing, ICCP 2017, Cluj-Napoca, Romania, September 7-9, 2017, 2017, pp. 231–238. doi:10.1109/ICCP.2017.8117009

[2] H. Wang, Q. Zhang, J. Wu, S. Pan, Y. Chen, Time series feature learning with labeled and unlabeled data, Pattern Recognit. 89 (2019) 55–66. doi:10.1016/j.patcog.2018.12.026

[3] I. Oregi, A. Pérez, J. D. Ser, J. A. Lozano, On-line elastic similarity measures for time series, Pattern Recognit. 88 (2019) 506–517. doi:10.1016/j.patcog.2018.12.007

[4] A. J. Bagnall, J. Lines, A. Bostrom, J. Large, E. J. Keogh, The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Data Min. Knowl. Discov. 31 (3) (2017) 606–660. doi:10.1007/s10618-016-0483-9.

[5] J. Sun, Y. Yang, Y. Liu, C. Chen, W. Rao, Y. Bai, Univariate time series classification using information geometry, Pattern Recognit. 95 (2019) 24–35.

[6] A. Abanda, U. Mori, J. A. Lozano, A review on distance based time series classification, Data Min. Knowl. Discov. 33 (2) (2019) 378–412. doi:10.1007/s10618-018-0596-4.

[7] K. S. Tuncel, M. G. Baydogan, Autoregressive forests for multivariate time series modeling, Pattern Recognit. 73 (2018) 202–215. doi:10.1016/j.patcog.2017.08.016.

[8] N. D. Cilia, C. D. Stefano, F. Fontanella, A. S. di Freca, A ranking-based feature selection approach for handwritten character recognition, Pattern Recognit. Lett. 121 (2019) 77–86. doi:10.1016/j.patrec.2018.04.007.

[9] S. Sharmin, M. Shoyaib, A. A. Ali, M. A. H. Khan, O. Chae, Simultaneous feature selection and discretization based on mutual information, Pattern Recognit. 91 (2019) 162–174. doi:10.1016/j.patcog.2019.02.016.

[10] G. Brown, A. C. Pocock, M. Zhao, M. Luján, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.

[11] J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information, Neural Comput. Appl. 24 (1) (2014) 175–186. doi:10.1007/s00521-013-1368-0.

[12] B. H. Nguyen, B. Xue, P. Andreae, Mutual information estimation for filter based feature selection using particle swarm optimization, in: Applications of Evolutionary Computation - 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 - April 1, 2016, Proceedings, Part I, 2016, pp. 719–736. doi:10.1007/978-3-319-31204-0_46.

[13] B. H. Nguyen, B. Xue, P. Andreae, Mutual information for feature selection: estimation or counting?, Evol. Intell. 9 (3) (2016) 95–110. doi:10.1007/s12065-016-0143-4.

[14] S. Gupta, A. D. Dileep, T. A. Gonsalves, A joint feature selection framework for multivariate resource usage prediction in cloud servers using stability and prediction performance, J. Supercomput. 74 (11) (2018) 6033–6068. doi:10.1007/s11227-018-2510-7.

[15] A. Motrenko, V. V. Strijov, Multi-way feature selection for ecog-based brain-computer interface, Expert Syst. Appl. 114 (2018) 402–413. doi:10.1016/j.eswa.2018.06.054.

[16] Z. Karevan, J. A. K. Suykens, Transductive feature selection using clustering-based sample entropy for temperature prediction in weather forecasting, Entropy 20 (4) (2018) 264. doi:10.3390/e20040264.

[17] T. Liu, H. Wei, K. Zhang, W. Guo, Mutual information based feature selection for multivariate time series forecasting, in: 35th Chinese Control Conference, CCC 2016, Chengdu, China, July 27-29, 2016, 2016, pp. 7110–7114. doi:10.1109/ChiCC.2016.7554480.

[18] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (2004) 066138. doi:10.1103/PhysRevE.69.066138.

[19] A. Gonzalez-Vidal, F. Jimenez, A. F. Gomez-Skarmeta, A methodology for energy multivariate time series forecasting in smart buildings based on feature selection, Energy and Buildings 196 (2019) 71–82. doi:10.1016/j.enbuild.2019.05.021.

[20] S. Cheema, T. Henne, U. Koeckemann, E. Prassler, Applicability of feature selection on multivariate time series data for robotic discovery, in: 3rd International Conference on Advanced Computer Theory and Engineering, ICACTE 2010, Chengdu, China, Aug. 20-22, 2010, Vol. 2, 2010, pp. 592–597. doi:10.1109/ICACTE.2010.5579484.

[21] B. Chakraborty, Feature selection and classification techniques for multivariate time series, in: 2nd International Conference on Innovative Computing, Information and Control,ICICIC 2007, Kumamoto, Japan, Sep 5-7, 2007, 2007, pp. 42–42. doi:10.1109/ICICIC.2007.309.

[22] G. He, W. Zhao, X. Xia, R. Peng, X. Wu, An ensemble of shapelet-based classifiers on inter-class and intra-class imbalanced multivariate time series at the early stage, Soft Comput. 23 (2018) 6097–6114. doi:10.1007/s00500-018-3261-3.

[23] A. Jovic, F. Jovic, Classification of cardiac arrhythmias based on alphabet entropy of heart rate variability time series, Biomed. Signal Proc. and Control 31 (2017) 217–230. doi:10.1016/j.bspc.2016.08.010.

[24] M. Han, W. Ren, X. Liu, Joint mutual information-based input variable selection for multivariate time series modeling, Eng. Appl. of AI 37 (2015) 250–257. doi:10.1016/j.engappai.2014.08.011.

[25] L. Cui, Y. Jiao, L. Bai, L. Rossi, E. R. Hancock, Adaptive feature selection based on the most informative graph-based features, in: International Workshop on Graph-Based Representations in Pattern Recognition, Springer, 2017, pp. 276–287. doi:10.1007/978-3-319-58961-9_25.

[26] L. Fang, H. Zhao, P. Wang, M. Yu, J. Yan, W. Cheng, P. Chenoelho, Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data, Biomed. Signal Proc. and Control 21 (2015) 82–89. doi:10.1016/j.bspc.2015.05.011.

[27] H. Yoon, K. Yang, C. Shahabi, Feature subset selection and feature ranking for multivariate time series, IEEE Trans. Knowl. Data Eng. 17 (9) (2005) 1186–1198. doi:10.1109/TKDE.2005.144.

[28] A. K. Shekar, M. Pappik, P. I. Sánchez, E. Müller, Selection of relevant and non-redundant multivariate ordinal patterns for time series classification, in: Discovery Science - 21st International Conference, DS 2018, Limassol, Cyprus, October 29-31, 2018, Proceedings, 2018, pp. 224–240. doi:10.1007/978-3-030-01771-2_15.

[29] A. Bondu, D. Gay, V. Lemaire, M. Boullé, E. Cervenka, FEARS: a feature and representation selection approach for time series classification, in: W. S. Lee, T. Suzuki (Eds.), Proceedings of The 11th Asian Conference on Machine Learning, ACML 2019, 17-19 November 2019, Nagoya, Japan, Vol. 101 of Proceedings of Machine Learning Research, PMLR, Nagoya, Japan, 2019, pp. 379–394.

[30] S. Gudmundsson, T. P. Runarsson, S. Sigurdsson, Test–retest reliability and feature selection in physiological time series classification, Comput. Meth. Prog. Bio. 105 (1) (2012) 50–60. doi:10.1016/j.cmpb.2010.08.005.

[31] A. Saikhu, A. Arifin, C. Fatichah, Correlation and symmetrical uncertainty-based feature selection for multivariate time series classification, International Journal of Intelligent Engineering and Systems 12 (3) (2019) 129–137.

[32] C. E. Shannon, A mathematical theory of communication, Mob. Comput. Commun. Rev. 5 (1) (2001) 3–55. doi:10.1145/584091.584093.

[33] T. Blumentritt, F. Schmid, Mutual information as a measure of multivariate association: analytical properties and statistical estimation, J. Stat. Comput. Sim. 82 (9) (2012) 1257–1274. doi:10.1080/00949655.2011.575782.

[34] M. Vollmer, K. Böhm, Iterative estimation of mutual information with error bounds, in: Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019, 2019, pp. 73–84. doi:10.5441/002/edbt.2019.08.

[35] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, S. Aigrain, Gaussian processes for time-series modelling, Philos. T. Roy. Soc. A 371 (2013) 20110550. doi:10.1098/rsta.2011.0550.

[36] N. H. H. Salah H. Abid, The entropy of arma (p,q) process, Am. J. Math. Stat. 4 (1) (2014) 12–20. doi:10.5923/j.ajms.20140401.03.

[37] C. R. Baker, Mutual information for gaussian processes, SIAM J. Appl. Math. 19 (2) (1970) 451–458. doi:10.1137/0119044.

[38] B. C. Ross, Mutual information between discrete and continuous data sets, PLoS One 9 (2) (2014) 1–5. doi:10.1371/journal.pone.0087357.

[39] A. Bulinski, A. Kozhevin, Statistical estimation of conditional shannon entropy, ESAIM: Probability and Statistics 23 (2019) 350–386. doi:10.1051/ps/2018026.

[40] F. Coelho, A. de Pádua Braga, M. Verleysen, A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems, Int. J. Comput. Intell. Syst. 9 (4) (2016) 726–733. doi:10.1080/18756891.2016.1204120

[41] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, Y. Zhou, A feature subset selection algorithm automatic recommendation method, J. Artif. Intell. Res. 47 (2013) 1–34. doi:10.1613/jair.3831.

[42] A. J. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, E. J. Keogh, The UEA multivariate time series classification archive, 2018, CoRR abs/1811.00075 (2018). arXiv:1811.00075.

[43] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, E. J. Keogh, Generalizing DTW to the multi-dimensional case requires an adaptive approach, Data Min. Knowl. Discov. 31 (1) (2017) 1–31. doi:10.1007/s10618-016-0455-0.

[44] I. Oregi, J. D. Ser, A. Pérez, J. A. Lozano, Nature-inspired approaches for distance metric learning in multivariate time series classification, in: 2017 IEEE Congress on Evolutionary Computation, CEC 2017, Donostia, San Sebastián, Spain, June 5-8, 2017, 2017, pp. 1992–1998. doi:10.1109/CEC.2017.7969545.