Application of functional kernel hypothesis testing for channel selection in time series classification

Huang, Q.; Bäck, T.H.W.; Stein, N. van

Application of Functional Kernel Hypothesis Testing for Channel Selection in Time Series Classification

Qi Huang
LIACS, Leiden University
Leiden, The Netherlands
0009-0007-4989-135X

Thomas Bäck
LIACS, Leiden University
Leiden, The Netherlands
0000-0001-6768-1478

Niki van Stein
LIACS, Leiden University
Leiden, The Netherlands
0000-0002-0013-7969

2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00111

***Abstract*—Multi-variate time series classification tasks are prevalent in various real-world engineering domains, including but not limited to activity recognition and anomaly detection. However, due to the abundance of sensors available, selecting the appropriate channels for successful classification can be a daunting task. In this study, we propose to use a two-sample hypothesis test, to determine the relevance of channels in time series classification tasks. We illustrate the industrial usecase that motivated this algorithm and validate our approach on open-source benchmarks. The proposed method has the potential to address the challenge of channel selection in multi-variate time series classification tasks and can significantly impact various real-world engineering applications.**

***Index Terms*—Channel Selection, Feature Selection, Time Series Classification**

I. Introduction

Let $X \sim \mathbb{P}_X$ be a stochastic process (channel), where $x_t = X(t)$ is its realization on a time point t and we let $X_I = [x_1, x_2, \dots, x_n]$ to represent the segment of X observed over a time interval I with n measurement points (t_i). Additionally, we define time series T, a segment-collection of X , as $T = \{T_1, T_2, \dots, T_k\}$, where $T_j = X_{I_j}$. Time series classification requires building a machine learning model to predict the class labels of unseen time series based on k historical data-label observations T and labels L. Assuming for each time interval I, we observe $M \geq 2$ stochastic processes simultaneously, i.e., $T_j = (X_{1,I_j}, \dots, X_{M,I_j})$, where X_{i,I_j} is the segment of channel i on interval I_j. Then the problem becomes multivariate or multi-channel time series classification.

When dealing with classification tasks in complex industrial environments, one of the challenges is the large number of sensors involved. Each sensor generates a sequence of outputs (a channel). It is often unclear which sensors have a direct or indirect relation with the classification target, posing difficulties in selecting the relevant channels for classification. To provide a concrete example, we consider the industrial use-case of classifying the activity of the world's largest crane vessel, the Sleipnir, owned by the Heerema company. This vessel is equipped with over $m = 6000$ sensors, measuring various subsystems such as thrusters, ballast pumps, cranes, and engines. Depending on the activity to be classified, many of these channels are potentially relevant, and selecting a feasible subset of channels is a challenging and time-consuming task. Hence, this research aims to handle the following scenario: **Given a TSC task on time series T with M channels (sensors), find the most $P \leq M$ relevant channels for predicting the segment-wise labels** L. For a robust TSC classifier f, a train set D_{tr}, a test set D_{te}, and a performance metric $P(D_{te})$ where a larger value means better performance. Herewith, we define a channel $X_{i,*}$ is *truly relevant* if it satisfies the following criteria:

- Fit f exclusively on $X_{i,train}, L$ and evaluate performance on the test data $D_{i,test}$. Then $P(D_{te}) - P(D_{i,te}) \leq \epsilon$.
- Remove $X_{i,*}$ from D_{tr} and D_{te}, and let the new train and test sets be D^i_{tr} and D^i_{te}. Train f on D^i_{tr} and obtain its performance $P(D^i_{te})$, then $P(D_{te}) - P(D^i_{te}) \geq \delta$,

where both ϵ and $\delta \geq 0$ are scenario-dependent thresholds.

It is possible to determine the relevance by exhaustively doing leave-one-channel-out runs. However, this violates the principle of channel selection, namely, saving computational costs. Thus, we propose a methodology that can quantify and estimate this relevance based on a functional kernel-based two-sample test [1] to determine the relevance of channels for a given TSC task. Due to the confidentiality of our industrial data, we verify our proposed method on several open-source benchmarks and demonstrate the effectiveness and limitations of the proposed approach.

II. Related Work

Although various feature selection methods exist for classical tabular machine learning, there is a notable research gap in terms of selecting channels for multi-channel time series tasks. Some existing methods employ static feature extraction to determine the relevant channels and compute inner-correlations, such as entropy, to rank the performance of different learning algorithms on different feature subsets or employ a hybrid of these two methods [2], [3]. Another approach is to study the correlations between channels, including joint correlations, using Pearson's correlation [4], [5]. However, we argue that Pearson's correlation does not account for the time-dependency of channels, which makes it imperfect for time-series analysis.

This publication is part of the project XAIPre (with project number 19455) of the research program Smart Industry 2020 which is (partly) financed by the Dutch Research Council (NWO).

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00111

III. Methodology

Time space to function space: It is assumed that each univariate time series is a function supported over a vector space, and can be represented by a finite vector [6].

Functional kernel hypothesis test: A functional kernel two-sample test aims to test the distributional equivalence between two functions supported by the same vector space. The core idea is to use a reproducing kernel defined on functional space to construct and test the metric discrepancies between the mean embedding of two functions [1].

Based on these preliminaries, for given k time intervals $I = \{I_1, \ldots, I_k\}$, we observe a univariate time series channel $X = \{X_{I_1}, \ldots, X_{I_k}\}$ and its element-wise labels $L = \{L_1, \ldots, L_k\}$, i.e., pairs of (X_{I_j}, L_j). Additionally, we assume the time difference (start to end) for all time intervals is the same but the sampling strategy for measurement points in each I_j can be different. Consequently, it is feasible for us to consider that all X_{I_j} are defined in the same space but with different observation grids. Moreover, let the functional kernel two-sample test algorithm be $g(F_p, F_q; \kappa)$, where $F_p \sim \mathbb{P}$ and $F_q \sim \mathbb{Q}$ are functional distributions and κ is the functional reproducing kernel. Then the g-based relevance of X w.r.t L is determined as follows:

1) Suppose there are m unique labels in L, namely, $A = \{A_1, \ldots, A_m\}$ and note that $X \to A$ is surjective-only. For each unique $A_i \in A$, we find the two subsets $X_i = \{X_{I_j} \mid L_j = A_i,\ \forall j \in [1, k]\}$ and $X_i^* = \{X \setminus X_i\}$.
2) For each pair of (X_i, X_i^*), we determine their two-sample-test-based test power, namely, $t_i = g(X_i, X_i^*; \kappa)$. It is obvious that here we consider X_i and X_i^* to be two different functions supported by I.
3) The relevance score of X w.r.t L bounded by κ, namely, $R(X \mid L, \kappa)$ can then be written as:
$$R(X \mid L, \kappa) = \sum\nolimits_{i=1}^{m} \frac{t_i}{m}$$
4) Suppose we have C valid functional reproducing kernels, i.e., $\Omega = \{\kappa_1, \ldots, \kappa_C\}$. We do procedures 1 to 3 for C times, each time using a unique $\kappa_j \in \Omega$. Then the final channel relevance of X w.r.t L is defined as:
$$R(X \mid L) = \max_{j=1}^{C} R(X \mid L, \kappa_j)$$

It is beneficial to use multiple kernels since each kernel is limited by its form, and can only capture certain types of hypothesis classes of the functions[1].

IV. Experiments

Setup: Recall the initial goal of finding the top $P \leq M$ relevant channels for TSC task. For efficiency, we generalize *relevance* to multi-channels by considering or removing a batch of channels that fall into a certain range of relevance. To verify the soundness of our proposed channel relevance, we perform benchmarking on the well-established UCR time series suites [7]. The experiments are organized as follows:

1) Given a m-channels multi-variate TSC dataset ($T = \{X_{1,*}, \ldots, X_{m,*}\}$). For each channel $X_{i,*}$, we compute its channel relevance score $R(X_{i,*}|L)$ based on the aforementioned algorithm.
2) Grouping channels by their channel relevance, e.g., $S_{[a,b]} = \{X_{j,*} \mid R(X_{j,*}|L) \in [a, b],\ \forall j \in [1, \ldots, m]\}$.
3) Training and testing the performance of a classifier on each of the $S_{[a,b]}$.

For each of the used datasets, the classifier which yields the best results w.r.t the extensive review by Ruiz et al. [8] is used for benchmarking, where each classifier has been independently run 10 times on the dataset.

In addition to the benchmark settings, we outline the five kernels that are utilized in this experiment. Suppose we are distinguishing between two functional distributions $F_p \sim \mathbb{P}$ and $F_q \sim \mathbb{Q}$ supported over $\mathbb{R}^d$, the kernel can be written as:

$$\kappa(F_p, F_q) = e^{-\frac{\|T(F_p) - T(F_q)\|^2}{2\eta^2}},$$

where η is intrinsically the bandwidth of the Gaussian kernel, T is a Borel measurable, continuous, and injective transformation that is capable of mapping both functions into a real and separable Hilbert space. In this study, we consider the following five kernels from [1]:

- Standard: $T(a) = a$, the original Gaussian RBF kernel
- Cosine (Cos): point-wise cosine transformation
- Squaring feature expansion (Sqr): summing up two kernels. One with $T(a) = a$ and another one with $T(a) = a^2$
- FPCA: T is functional PCA.
- Covariance (Cov): it is different from other kernels, the outcome is the inner product between F_p and F_q.

Results: Four symbolic results obtained on four datasets from the UCR multivariate suite [9] are shown in Table I,II,III and IV.

TABLE I
Results obtained on *RacketSports*

<table>
<tr><th></th><th colspan="3">Range of channel relevance</th></tr>
<tr><th></th><th>(0.4,0.6]</th><th>(0.6,0.8]</th><th>(0.8,1.0]</th></tr>
<tr><td rowspan="4">F1</td><td>0.80</td><td>0.82</td><td>0.83</td></tr>
<tr><td colspan="2">0.85</td><td>0.83</td></tr>
<tr><td>0.80</td><td colspan="2"><b>0.88</b></td></tr>
<tr><td colspan="3">0.86</td></tr>
<tr><td>N</td><td>2</td><td>2</td><td>2</td></tr>
</table>

TABLE II
Results obtained on *HandMovementDirection*

<table>
<tr><th></th><th colspan="4">Range of channel relevance</th></tr>
<tr><th></th><th>[0,0.1]</th><th>(0.1,0.2]</th><th>(0.2,0.3]</th><th>(0.3,0.4]</th></tr>
<tr><td rowspan="5">F1</td><td>0.39</td><td>0.45</td><td>0.46</td><td>0.48</td></tr>
<tr><td colspan="2">0.43</td><td colspan="2"><b>0.55</b></td></tr>
<tr><td colspan="3">0.47</td><td>0.48</td></tr>
<tr><td>0.39</td><td colspan="3">0.52</td></tr>
<tr><td colspan="4">0.52</td></tr>
<tr><td>N</td><td>4</td><td>3</td><td>2</td><td>1</td></tr>
</table>

[1] However, considering two channels both with high relevancy but are based on fundamentally different kernels, then multiple kernels can be misleading since a classifier may fail to process two intrinsically different function classes.

We first give an example of how to read the tables, ”F1” means the (weighted) $f1$ scores (averaged over 10 independent runs) and N means the number of channels that falls into the relevance range (denoted at the top of each column). The two results in the second row of Table II shows the $f1$ scores 0.43 and 0.55 obtained by the classifier on channels whose relevance are between $[0, 0.2]$ and $(0.2, 0.4]$, respectively. The darker the color of a cell is, the better the performance. The best results per data set are shown in **bold** font.

It can be seen from Table I, by leaving out the two channels with *lower relevance*, the classifier yields better performance than using all channels. Similar and more encouraging results can be seen in Table II, where only relying on the top 3 relevant channels outperforms using all 10 channels.

TABLE III
RESULTS OBTAINED ON *Heartbeat*

<table>
<tr><th></th><th colspan="7">Range of channel relevance</th></tr>
<tr><th></th><th>[0,0.15]</th><th>(0.15,0.3]</th><th>[0.3,0.45]</th><th>[0.45,0.6]</th><th>(0.6,0.75]</th><th>(0.75,0.9]</th><th>(0.9,1.0]</th></tr>
<tr><td rowspan="8">F1</td><td>0.70</td><td>0.71</td><td>0.70</td><td>0.73</td><td>0.76</td><td>0.69</td><td>0.76</td></tr>
<tr><td>0.70</td><td colspan="6"><b>0.77</b></td></tr>
<tr><td colspan="2">0.71</td><td colspan="5"><b>0.77</b></td></tr>
<tr><td colspan="3">0.70</td><td colspan="4">0.76</td></tr>
<tr><td colspan="4">0.72</td><td colspan="3"><b>0.77</b></td></tr>
<tr><td colspan="5">0.74</td><td colspan="2">0.76</td></tr>
<tr><td colspan="6">0.74</td><td>0.76</td></tr>
<tr><td colspan="7"><b>0.77</b></td></tr>
<tr><td>N</td><td>1</td><td>2</td><td>2</td><td>5</td><td>10</td><td>10</td><td>31</td></tr>
</table>

TABLE IV
RESULTS OBTAINED ON *NATOPS*

<table>
<tr><th></th><th colspan="7">Range of channel relevance</th></tr>
<tr><th></th><th>(0.3,0.4]</th><th>(0.4,0.5]</th><th>(0.5,0.6]</th><th>(0.6,0.7]</th><th>(0.7,0.8]</th><th>(0.8,0.9]</th><th>(0.9,1.0]</th></tr>
<tr><td rowspan="8">F1</td><td>0.69</td><td>0.67</td><td>0.81</td><td>0.78</td><td>0.75</td><td>0.60</td><td>0.88</td></tr>
<tr><td>0.69</td><td colspan="6"><b>0.91</b></td></tr>
<tr><td colspan="2">0.75</td><td colspan="5"><b>0.91</b></td></tr>
<tr><td colspan="3">0.82</td><td colspan="4"><b>0.91</b></td></tr>
<tr><td colspan="4">0.87</td><td colspan="3">0.89</td></tr>
<tr><td colspan="5">0.89</td><td colspan="2">0.87</td></tr>
<tr><td colspan="6"><b>0.91</b></td><td>0.88</td></tr>
<tr><td colspan="7"><b>0.91</b></td></tr>
<tr><td>N</td><td>1</td><td>4</td><td>4</td><td>3</td><td>2</td><td>2</td><td>8</td></tr>
</table>

In Table III, by leaving out 10 channels the classifier is still performing as good as including all channels. One could even opt to leave out the 30 least informative channels and only lose 0.01 on the $f1$ score. In principle, the values on the right side should be higher than the ones on the left side, as these include the most relevant channels, but this is not guaranteed since the proposed channel relevance is based on correlation instead of causality and the inter-channel information shall be taken into account. From Table IV one can observe that leaving out the 8 most relevant channels leads to the same $f1$ score as leaving out the 9 least relevant channels. Looking at the channel relevance scores of the NATOPS benchmark, we can observe that actually all channels show some relevance and therefore this does make sense.

Lastly, Table V shows the number of times the relevance of a channel is determined by a certain type of kernel. Looking at the **bold** numbers, it is not hard to see that the squaring function expansion is the most confident kernel on three out of four problems (Heartbeat (HB), RacketSports (RS), and NATOPS), which suggests the original signals could possess strong polynomial inner-correlations. On the contrary, it can be assumed that there exist more sine-cosine correlations within the data of HandMovementDirection (HMD).

TABLE V
NUMBER OF TIMES THAT A KERNEL OBTAINS THE HIGHEST RELEVANCE

	Standard	Cos	Sqr	Cov	FPCA	Total
HMD	1	**7**	0	0	2	10
HB	0	0	**61**	0	0	61
RS	0	0	**4**	2	0	6
NATOPS	1	1	**19**	3	0	24

V. CONCLUSIONS AND OUTLOOK

A novel functional kernel-based two-sample test approach is proposed to determine the channel relevance in multivariate time series classification tasks. The effectiveness of the approach is validated on a wide set of benchmark TSC problems. We show that the two-sample test methodology, combining different static kernels, works very well in order to reduce the number of channels and even increase classification accuracy (f1 score) in some cases. Although the experimental results show that one kernel is promising most of the time, choosing the best static functional kernel as well as selecting hyper-parameters can however still pose a challenge for real-world applications, and the inter-channel relevance is not taken into account using the proposed approach. For future research directions, a learned kernel could be interesting to solve the kernel selection problem.

REFERENCES

[1] G. Wynne and A. B. Duncan, “A Kernel Two-Sample Test for Functional Data,” *Journal of Machine Learning Research*, vol. 23, no. 73, pp. 1–51, 2022. [Online]. Available: http://jmlr.org/papers/v23/20-1180.html

[2] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, “A review of channel selection algorithms for EEG signal processing,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 66, Aug. 2015. [Online]. Available: https://doi.org/10.1186/s13634-015-0251-9

[3] M. Z. Baig, N. Aslam, and H. P. H. Shum, “Filtering techniques for channel selection in motor imagery EEG applications: a survey,” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1207–1232, Feb. 2020. [Online]. Available: https://doi.org/10.1007/s10462-019-09694-8

[4] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, “Correlation-based channel selection and regularized feature optimization for MI-based BCI,” *Neural Networks*, vol. 118, pp. 262–270, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019301960

[5] D. D. Chakladar and S. Chakraborty, “EEG based emotion classification using “Correlation Based Subset Selection”,” *Biologically Inspired Cognitive Architectures*, vol. 24, pp. 98–106, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212683X18300227

[6] L. Horváth and P. Kokoszka, “Inference for Functional Data with Applications,” ser. Springer Series in Statistics, vol. 200. New York, NY: Springer New York, 2012. [Online]. Available: https://link.springer.com/10.1007/978-1-4614-3655-3

[7] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, “The UCR time series archive,” *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, Nov. 2019.

[8] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances,” *Data Mining and Knowledge Discovery*, vol. 35, no. 2, pp. 401–449, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s10618-020-00727-3

[9] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, and Hexagon-ML, “The ucr time series classification archive,” October 2018, https://www.cs.ucr.edu/∼eamonn/time_series_data_2018/.