EAST experimentation: additional material

Xavier Renard[1,3], Maria Rifqi[2], Gabriel Fricout[3] and Marcin Detyniecki[1,4]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6, Paris, France.
[2]Université Panthéon Assas, Univ Paris 02, LEMMA, Paris, France.
[3]Arcelormittal Research, Maizières-lès-Metz, France.
[4]Polish Academy of Sciences, IBS PAN, Warsaw, Poland.

Precision on the methodology & the statistical assessment of the results

A strict evaluation protocol is required to assess the EAST representation and the random shapelets because they contain a random generation step. We rely on the evaluation protocol proposed in [1] for a proper way to analyze the performances of randomized algorithms. Each single test of the the EAST representation and the random shapelets is reproduced 10 times to evaluate the variability. 10 times is the minimum recommended in [1], we didn't perform more because the current number of configurations tested required weeks of calculations on a cluster.

The raw classification accuracies are presented table 1 for $q = 2000$. For randomized algorithms, the mean accuracy for each configuration and dataset is shown together with the standard deviation. The results for other values of q are available in the folder *results* of the website of this work [3].

To evaluate the statistical significance of the performances between two randomized algorithms we use a non parametric paired Wilcoxon test to assess if the differences between their classification accuracies are centered around 0. The null hypothesis H_0 of this test is the absence of difference. The $p-value$ is conserved to get the probability to reject H_0 while the performances are actually identical (ie. *type I error*). We set up $\alpha = 0.05$ as the limit for the *type I error* in order to determine if the differences are statistically significant or not. To compare a randomized algorithm with a deterministic one (here the shapelet ensemble), we use a similar procedure with a non parametric one sample Wilcoxon test.

We evaluate the RFE+SVM, RLR+SVM, RLR+RF, RF, RSHPT and SHPT approaches. All the randomized algorithms are evaluated with 6 values for q. Thus we evaluate the performances of 31 algorithms and configurations. In order to aggregate the results over all the datasets we use the procedure recommended in [1]. At the beginning of the procedure every configuration is given a score set up to 0. For each Wilcoxon test between two algorithms, if the difference is significant the score of the best performing algorithm is increased by 1, the other is decreased by 1. The result is the ranking presented Fig. 3 of the paper, from the best performing overall the tested algorithms (with the highest score) to the worst performing (with the lowest score).

Finally, we compute the critical difference with the Nemenyi test with $\alpha = 0.05$ for the average ranks of the approaches on the datasets tested. Fig. 2 of the paper shows the critical difference with the average ranks for all the approaches. Connected approaches don't have significantly different classification performances according to the performed Nemenyi test [2].

Raw classification results

	RLR+SVM	RLR+RF	RF	RFE+SVM	RSHPT	SHPT
50words	70.2% ± 0.4	61.7% ± 1.4	61.0% ± 1.7	69.9% ± 0.5	12.5% ± 0.0	71.9%
Adiac	43.0% ± 1.8	66.5% ± 1.6	66.6% ± 2.4	40.9% ± 4.8	27.8% ± 3.7	56.5%
Beef	61.0% ± 2.7	61.3% ± 6.9	59.7% ± 5.8	61.0% ± 8.5	37.7% ± 2.7	83.3%
CBF	99.6% ± 0.2	97.5% ± 1.7	96.3% ± 2.8	99.4% ± 0.3	33.1% ± 0.0	99.7%
Car	73.8% ± 1.6	73.7% ± 3.8	73.2% ± 3.9	70.7% ± 1.4	27.7% ± 4.7	73.3%
ChlorineConcentration	56.5% ± 0.2	60.0% ± 0.7	60.0% ± 1.0	56.0% ± 0.1	53.8% ± 0.4	70.0%
CinCECGtorso	82.3% ± 1.8	81.8% ± 2.6	79.0% ± 3.6	73.6% ± 2.6	24.8% ± 0.0	84.6%
Coffee	92.9% ± 0.0	91.1% ± 2.5	90.0% ± 1.5	94.3% ± 2.5	81.8% ± 3.9	100.0%
CricketX	72.6% ± 0.8	61.8% ± 1.9	61.1% ± 2.6	71.8% ± 1.0	6.7% ± 0.0	78.2%
CricketY	71.8% ± 0.9	61.9% ± 2.2	61.1% ± 2.2	72.8% ± 0.5	9.1% ± 3.0	76.4%
CricketZ	74.7% ± 1.0	63.1% ± 2.4	63.0% ± 2.3	74.1% ± 1.7	6.2% ± 0.0	77.2%
DiatomSizeReduction	90.4% ± 0.7	89.7% ± 5.2	88.3% ± 4.5	89.2% ± 2.2	87.7% ± 1.1	87.6%
ECGFiveDays	98.3% ± 1.5	98.7% ± 1.3	90.5% ± 8.1	98.0% ± 1.1	50.6% ± 0.3	99.9%
FISH	87.0% ± 1.6	85.3% ± 3.0	83.8% ± 1.5	85.9% ± 1.9	37.3% ± 8.9	97.7%
FaceAll	75.9% ± 0.4	72.4% ± 0.6	71.7% ± 0.8	76.9% ± 0.5	15.7% ± 2.7	73.7%
FaceFour	99.0% ± 0.8	96.0% ± 2.4	94.4% ± 1.9	95.2% ± 1.3	39.3% ± 8.8	94.3%
FacesUCR	89.9% ± 0.4	81.2% ± 1.6	82.0% ± 1.7	88.6% ± 1.1	14.3% ± 0.0	91.3%
GunPoint	73.3% ± 1.5	85.5% ± 2.5	91.4% ± 2.2	71.1% ± 0.5	72.5% ± 5.7	98.0%
Haptics	48.5% ± 1.3	44.4% ± 2.1	44.0% ± 2.0	47.0% ± 1.0	20.8% ± 0.0	47.7%
InlineSkate	28.4% ± 1.1	30.9% ± 1.8	30.7% ± 1.8	25.8% ± 0.9	17.8% ± 0.4	38.5%
ItalyPowerDemand	95.7% ± 0.2	93.7% ± 0.6	94.2% ± 0.8	92.7% ± 6.5	93.7% ± 0.7	95.2%
Lighting2	72.3% ± 2.1	74.1% ± 2.9	72.6% ± 4.0	72.3% ± 2.7	54.1% ± 0.0	65.6%
Lighting7	71.6% ± 1.7	69.5% ± 2.9	68.8% ± 2.6	69.6% ± 3.9	26.0% ± 0.0	74.0%
MALLAT	92.0% ± 1.1	94.3% ± 1.7	95.9% ± 2.1	86.6% ± 2.8	13.1% ± 0.3	94.0%
MedicalImages	71.4% ± 0.9	67.4% ± 1.5	67.3% ± 1.2	67.6% ± 2.1	51.6% ± 0.1	60.4%
MoteStrain	86.1% ± 0.6	82.9% ± 1.6	82.9% ± 2.7	77.0% ± 9.8	54.9% ± 0.6	91.5%
NonInvasiveFatalECGThorax1	87.4% ± 0.5	84.6% ± 0.8	84.6% ± 0.6	87.0% ± 2.2	32.5% ± 7.9	90.0%
NonInvasiveFatalECGThorax2	89.7% ± 0.3	88.7% ± 0.3	88.0% ± 0.5	90.3% ± 1.8	31.4% ± 7.8	90.3%
OSULeaf	79.8% ± 1.8	69.2% ± 2.3	66.7% ± 2.2	79.1% ± 1.3	18.2% ± 0.0	71.5%
OliveOil	87.0% ± 2.5	84.0% ± 4.9	85.0% ± 4.8	86.0% ± 2.1	40.0% ± 0.0	90.0%
SonyAIBORobotSurface	96.1% ± 0.7	94.3% ± 2.2	91.7% ± 3.4	79.3% ± 10.6	42.9% ± 0.0	93.3%
SonyAIBORobotSurfaceII	91.5% ± 1.1	86.5% ± 2.3	85.4% ± 2.0	88.3% ± 5.2	61.7% ± 0.0	88.5%
StarLightCurves	95.2% ± 0.4	95.8% ± 0.2	96.1% ± 0.2	96.0% ± 0.4	57.7% ± 0.0	97.6%
SwedishLeaf	88.7% ± 0.6	85.7% ± 0.9	86.1% ± 1.7	88.2% ± 0.7	36.0% ± 8.0	90.7%
Symbols	93.8% ± 0.9	83.7% ± 3.9	85.4% ± 5.5	93.3% ± 1.0	17.4% ± 0.0	88.6%
Trace	99.9% ± 0.3	99.8% ± 0.4	99.2% ± 0.8	90.5% ± 9.9	42.6% ± 6.9	98.0%
TwoLeadECG	93.0% ± 0.8	90.7% ± 1.4	89.1% ± 2.3	92.6% ± 1.4	75.8% ± 6.5	99.6%
TwoPatterns	99.9% ± 0.1	98.7% ± 0.2	98.4% ± 0.3	99.6% ± 0.2	25.9% ± 0.0	94.1%
WordSynonyms	60.9% ± 1.0	54.8% ± 1.5	52.6% ± 0.9	60.3% ± 1.1	21.9% ± 0.0	59.7%
syntheticcontrol	99.1% ± 0.3	97.7% ± 0.6	98.2% ± 0.5	99.0% ± 0.5	57.1% ± 7.0	98.3%
uWaveGestureLibraryX	81.2% ± 0.2	76.4% ± 0.6	76.2% ± 0.6	80.6% ± 0.3	12.1% ± 0.0	78.4%
uWaveGestureLibraryY	72.0% ± 0.2	67.6% ± 0.7	67.2% ± 0.6	72.1% ± 0.3	12.1% ± 0.0	69.7%
uWaveGestureLibraryZ	74.3% ± 0.4	71.0% ± 0.4	70.8% ± 0.5	74.2% ± 0.4	12.1% ± 0.0	72.7%
wafer	99.7% ± 0.1	99.3% ± 0.2	99.0% ± 0.1	99.7% ± 0.0	90.6% ± 0.4	99.8%
yoga	79.6% ± 1.0	77.6% ± 1.7	80.8% ± 1.2	81.0% ± 0.5	54.0% ± 0.4	80.5%

Table 1. Detail of the accuracies for the various approaches with the standard deviation for random-based approaches (with $q = 2000$).

References

1. Andrea Arcuri and Lionel Briand. A Hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing Verification and Reliability*, 24(3):219–250, 2014.
2. Janez Demsar. Statistical Comparisons of Classifiers overMultiple Data Sets. *Journal of Machine Learning Research*, pages 1–30, 2006.
3. Xavier Renard, Maria Rifqi, Gabriel Fricout, and Marcin Detyniecki. https://github.com/xrenard/EAST-Representation, 2016.