IASD - NoSQL Presentation : Time Series Distance Measures Myths

Alexandre OLECH

Method

Widely adopted beliefs about Distance Measures in the TS community, due to the inconclusive and unchallenged works by Ding et al. [2] [3]:

- z-normalization is the go-to normalization.
- ED is the go-to lock-step measure.
- Elastic measures are better than sliding measures.
- DTW is the best elastic measure.

Challenge these claims with a more comprehensive and rigorous setup :

- 71 distance measures (lock-step, sliding, elastic, kernel, embedding).
- 8 normalizations.
- 128 datasets (UCR archive).
- Statistical tests (Wilcoxon [7], Friedman [4], Nemenyi [5]).

Method

Algorithm 1: Experimental Evaluation Setup

foreach *Distance Measure* **do**
    foreach *Normalization* **do**
        Compute the average accuracy of 1NN over 128 datasets of the UCR archive;

Assess the significance of performance differences with statistical tests:

- Wilcoxon test (95% confidence level) [7] (better than the t-test according to Rice 2006 [6])
- Friedman test [4] + Nemenyi test (90% confidence level) [5]

Main Results of the article

- Normalization significantly affects performance. For some distances, z-score is not the best.
- ED is not the best lock-step distance.
- Elastic measures are not ALWAYS better than sliding measures.
- DTW is not the best elastic measure.
- Best results : elastic and kernel configurations
- Embedding methods are almost as good and provide a very promising runtime-to-accuracy trade-off.

Weak Points

I'm very convinced by the method and results. Just a few remarks to improve the reproducibility :

- Clarify the implications of the z-score pre-normalization in the UCR archive on the scope of the results.
- Upload a checkpoint of the UCR archive.
- Website for full results exploration not accessible (redirects to the code page).
- Clarifying the relevance of including datasets with 0.00% accuracy (only 4 datasets on 80).

Strong Points

- Comprehensive experimental setup (128 datasets, 71 measures, 8 normalizations).
- Not just reporting accuracies, but also their statistical significance !
- The courage to drop a *pavé dans la marre* vs. the community's widespread misconceptions.

Is the paper solving an ongoing research problem?

- Challenging widely spread beliefs and widely-unquestioned results of previous works (Ding et al. [2], Ding et al. [3]).
- More comprehensive experimental evaluation, backed by rigorous statistical tests, to an old problem.

Is the paper improving existing state-of-the-art performances of a task?

Yes : successfully identifies normalizations and distances statistically better than the SOTA

Is the paper opening new research directions and problems?

Yes :

- Re-opening the question of distances/normalizations for Time-Series.
- Evaluating the effect of adopting a greater variety of distance measures into Time Series projects.
- Very interesting runtime-to-accuracy of embedding methods. Opens promising directions for these methods.

Are the experimental results covering the claim of the paper?

Yes :

- A very comprehensive evaluation of 71 distance measures, 8 normalization methods, over 128 time-series datasets.
- Using the mainstream evaluation method (1NN accuracy).
- Backed by well-sourced statistical tests (Demsar 2006 [1], Rice 2006 [6]).

Are there missing related work?

Not to my knowledge. Authors correctly identified the main (limited) references considered on the question (works by Ding et al. [2] [3]).

Implementation used by the author

- Easily accessible Matlab/C/Java code.
- Good reference for distance measures.
- Small mismatch (possibly outdated summary ?) between data and summary in UCR archive.

My Implementation

Goal : lock-step distances (table 2 of the article).

Distance Measure	Scaling Method	Better	Average Accuracy	>	=	<
Minkowski (L_p-norm)	z-score	✔	0.7083	79	13	36
	MinMax	✔	0.7041	70	12	46
	UnitLength	✔	0.7083	79	13	36
	MeanNorm	✔	0.7082	81	10	37
	Tanh	✘	0.6941	60	7	61
Lorentzian	z-score	✔	0.7022	71	8	49
	MinMax	✔	0.7010	66	7	55
	UnitLength	✔	0.7024	76	9	43
	MeanNorm	✔	0.7061	75	9	44
	Tanh	✘	0.6950	63	9	56
Manhattan (L_1-norm)	z-score	✔	0.7017	76	11	41
	MinMax	✔	0.7017	66	11	51
	UnitLength	✔	0.7017	76	11	41
	MeanNorm	✔	0.7051	76	9	43
	Tanh	✘	0.6913	63	11	54
Avg L_1/L_∞	z-score	✔	0.7012	75	10	43
	MinMax	✔	0.7013	68	5	55
	UnitLength	✔	0.7012	75	10	43
	MeanNorm	✔	0.7046	76	9	43
	Tanh	✘	0.6911	60	13	55
DISSIM	z-score	✔	0.7013	78	6	44
	MinMax	✔	0.7016	66	8	54
	UnitLength	✔	0.7013	78	6	44
	MeanNorm	✔	0.7039	73	9	46
	Tanh	✘	0.6917	64	10	54
Jaccard	MinMax	✘	0.6955	66	12	50
	MeanNorm	✔	0.6939	76	19	33
ED (L_2-norm)	MinMax	✘	0.6947	69	13	46
	MeanNorm	✘	0.6896	67	11	50
Emanon4	MinMax	✔	0.7034	72	6	50
Soergel	MinMax	✔	0.7011	73	4	51
Clark	MinMax	✘	0.6986	73	4	51
Topsoe	MinMax	✘	0.6962	71	4	53
Chord	MinMax	✘	0.6934	64	8	56
ASD	MinMax	✘	0.6884	56	13	59
Canberra	MinMax	✘	0.6933	56	4	68
ED	**z-score**	-	**0.6863**	-	-	-

Did you succeed?

Yes, but some simplifications in my implementation :

- Only considers datasets which are both in the UCR folder and in the UCR data summary (small inconsistencies between the two).
- Only consider datasets in the UCR archive which have a rough cost estimate of $c = n_{test} * n_{train} * d < 1e8$ (4h30 runtime).
- Only considers datasets with fixed length.
- Results in 80 datasets, vs. 128 in the original article.
- Only accuracies, no statistical tests.

Are the results consistent with those in the paper?

Consistent with the paper (excluding Minkowski : requires tuning):

- ED and Jaccard outperformed by Lorentzian, Manhattan and Average L_1, L_{inf}.
- z-score not always the best.
- z-score is systematically outperformed by MeanNorm and UnitLength.
- Tanh normalization offers the poorest results.
- Excluding Tanh, the global range of accuracy values is very close.

Slightly different from the paper :

- Minkowski is not as competitive (doesn't really count : requires tuning and was not tuned).
- Tanh leads to lower accuracies (68% vs. 65-66% in the article).

Frame Title

Article (128 datasets)

Distance Measure	Scaling Method	Better	Average Accuracy	>	=	<
Minkowski (L_p-norm)	*z*-score	✔	0.7083	79	13	36
	MinMax	✔	0.7041	70	12	46
	UnitLength	✔	0.7083	79	13	36
	MeanNorm	✔	0.7082	81	10	37
	Tanh	✘	0.6941	60	7	61
Lorentzian	*z*-score	✔	0.7022	71	8	49
	MinMax	✔	0.7010	66	7	55
	UnitLength	✔	0.7024	76	9	43
	MeanNorm	✔	0.7061	75	9	44
	Tanh	✘	0.6950	63	9	56
Manhattan (L_1-norm)	*z*-score	✔	0.7017	76	11	41
	MinMax	✔	0.7017	66	11	51
	UnitLength	✔	0.7017	76	11	41
	MeanNorm	✔	0.7051	76	9	43
	Tanh	✘	0.6913	63	11	54
Avg L_1/L_∞	*z*-score	✔	0.7012	75	10	43
	MinMax	✔	0.7013	68	5	55
	UnitLength	✔	0.7012	75	10	43
	MeanNorm	✔	0.7046	76	9	43
	Tanh	✘	0.6911	60	13	55
DISSIM	*z*-score	✔	0.7013	78	6	44
	MinMax	✔	0.7016	66	8	54
	UnitLength	✔	0.7013	78	6	44
	MeanNorm	✔	0.7039	73	9	46
	Tanh	✘	0.6917	64	10	54
Jaccard	MinMax	✘	0.6955	66	12	50
	MeanNorm	✔	0.6939	76	19	33
ED (L_2-norm)	MinMax	✘	0.6947	69	13	46
	MeanNorm	✘	0.6896	67	11	50
Emanon4	MinMax	✔	0.7034	72	6	50
Soergel	MinMax	✔	0.7011	73	4	51
Clark	MinMax	✘	0.6986	73	4	51
Topsoe	MinMax	✘	0.6962	71	4	53
Chord	MinMax	✘	0.6934	64	8	56
ASD	MinMax	✘	0.6884	56	13	59
Canberra	MinMax	✘	0.6933	56	4	68
ED	***z*-score**	-	**0.6863**	-	-	-

Reproduction (80 datasets)

Distance Measure	Scaling Method	Average Accuracy
Minkowski	z-score	66.1757
Minkowski	MinMax	68.0005
Minkowski	UnitLength	67.0077
Minkowski	MeanNorm	67.3444
Minkowski	Tanh	64.1729
Lorentzian	z-score	69.8322
Lorentzian	MinMax	69.2759
Lorentzian	UnitLength	70.4878
Lorentzian	MeanNorm	70.2331
Lorentzian	Tanh	66.2805
Manhattan	z-score	69.6402
Manhattan	MinMax	69.3353
Manhattan	UnitLength	70.3871
Manhattan	MeanNorm	70.0615
Manhattan	Tanh	66.1142
Avg_l1_linf	z-score	69.4754
Avg_l1_linf	MinMax	69.3259
Avg_l1_linf	UnitLength	70.2137
Avg_l1_linf	MeanNorm	70.0042
Avg_l1_linf	Tanh	66.0858
Jaccard	z-score	67.9041
Jaccard	MinMax	68.7838
Jaccard	UnitLength	68.6311
Jaccard	MeanNorm	68.674
Jaccard	Tanh	64.9398
ED	z-score	67.896
ED	MinMax	68.8665
ED	UnitLength	68.6311
ED	MeanNorm	69.0077
ED	Tanh	65.1855

Are the results different, but the trends and the claims still hold?

Yes, the claims about Myths 1 & 2 completely hold :

- z-score is not always the best normalization.
- ED can be outperformed by other lock-step measure.

Final Verdict

- Comprehensive experimental setup.
- Rigorous statistical tests.
- Reproducible results.
- Great quality.
- Great contribution.

Based on these, I fully accept the article for publication.

References

[1] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *The Journal of Machine Learning Research* 7 (2006), pp. 1–30.

[2] Hui Ding et al. “Querying and mining of time series data: experimental comparison of representations and distance measures”. In: *Proceedings of the VLDB Endowment* 1.2 (2008), pp. 1542–1552.

[3] Rui Ding et al. “Yading: Fast clustering of large-scale time series data”. In: *Proceedings of the VLDB Endowment* 8.5 (2015), pp. 473–484.

[4] Milton Friedman. “The use of ranks to avoid the assumption of normality implicit in the analysis of variance”. In: *Journal of the American Statistical Association* 32 (1937), pp. 675–701.

[5] Peter Nemenyi. “Distribution-free Multiple Comparisons”. PhD thesis. Princeton University, 1963.

[6] John Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 2006.