Fast Time Series Classification with Random Symbolic Subsequences

Thach Le Nguyen – Georgiana Ifrim

University College Dublin, Ireland

HOST INSTITUTION

Agriculture and Food Development Authority

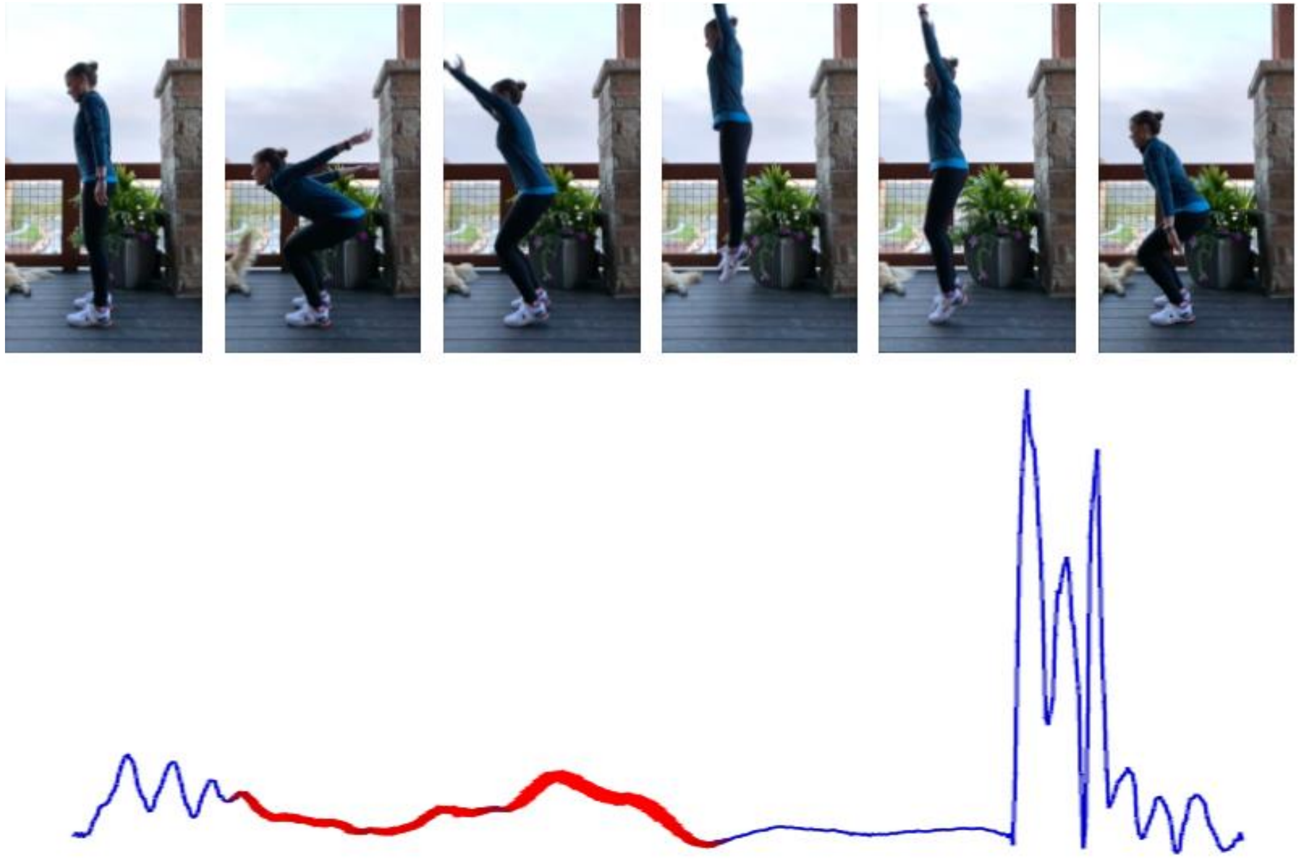
PARTNER INSTITUTIONS

FUNDED BY:

An Roinn Talmhaíochta, Bia agus Mara
Department of Agriculture, Food and the Marine

Digitalising Dairy

Time Series Classification (TSC)

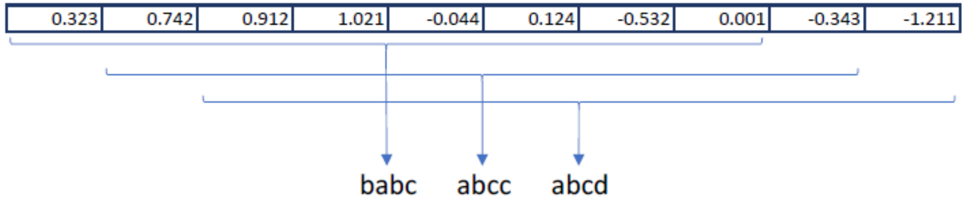


Time series classification is the problem of labelling unseen time series.

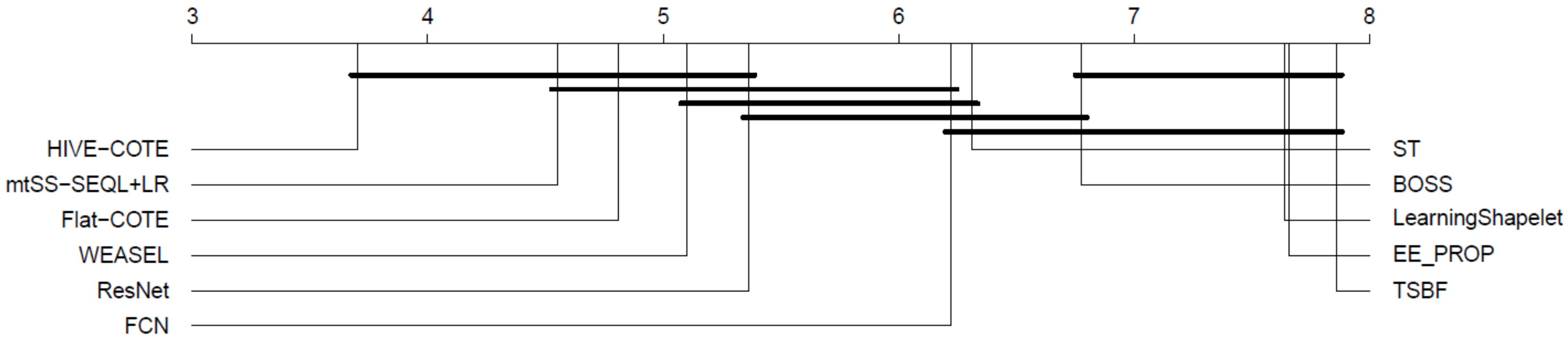
For example: Did the participant perform a successful countermovement jump ?

Motivation

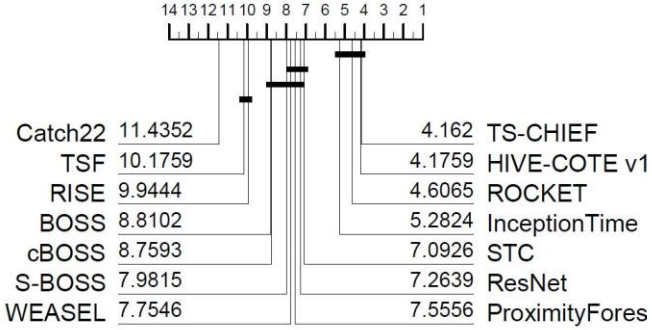
- SAX (2002), SFA (2012)



- BOSS(2015), WEASEL (2017)
- MrSEQL (2019)



- ROCKET (2020)
- COTE (2015) , HIVE-COTE (2018), TS-CHIEF (2020), InceptionTime (2020)



Motivation

- A challenge of symbolic time series classifiers is the large feature space.
- However, do we really need complex feature selection methods (e.g. in WEASEL and MrSEQL) ?
- Linear models appear to be still a good choice for time series classification.
- Fast and simple methods always have the advantage in practice.

Contributions

- We propose MrSQM, a symbolic time series classifier which builds on multiple symbolic representations, random sequence mining and a linear classifier.
- Extensive empirical study comparing accuracy and runtime of MrSQM to recent state-of-the-art.
- All our code and data is publicly available at https://github.com/mlgig/mrsqm

Methodology

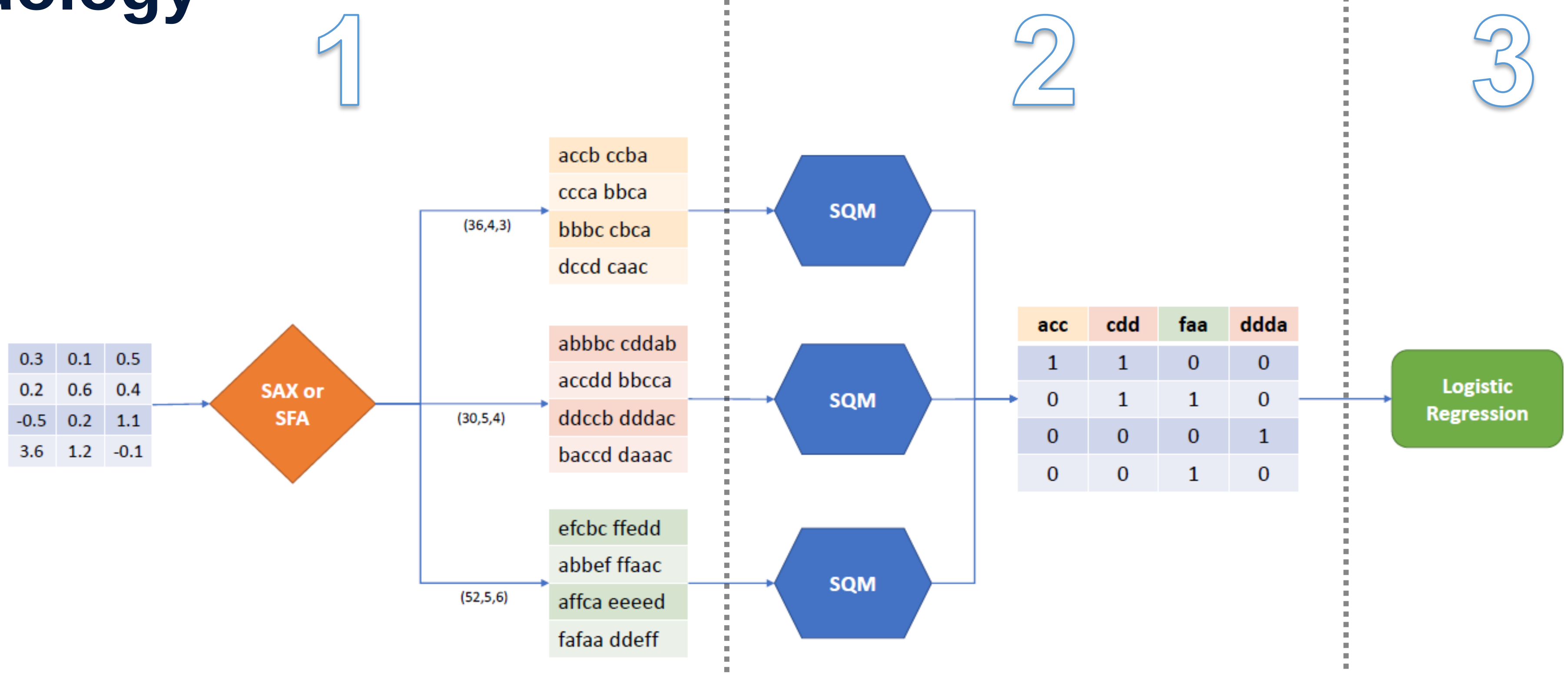


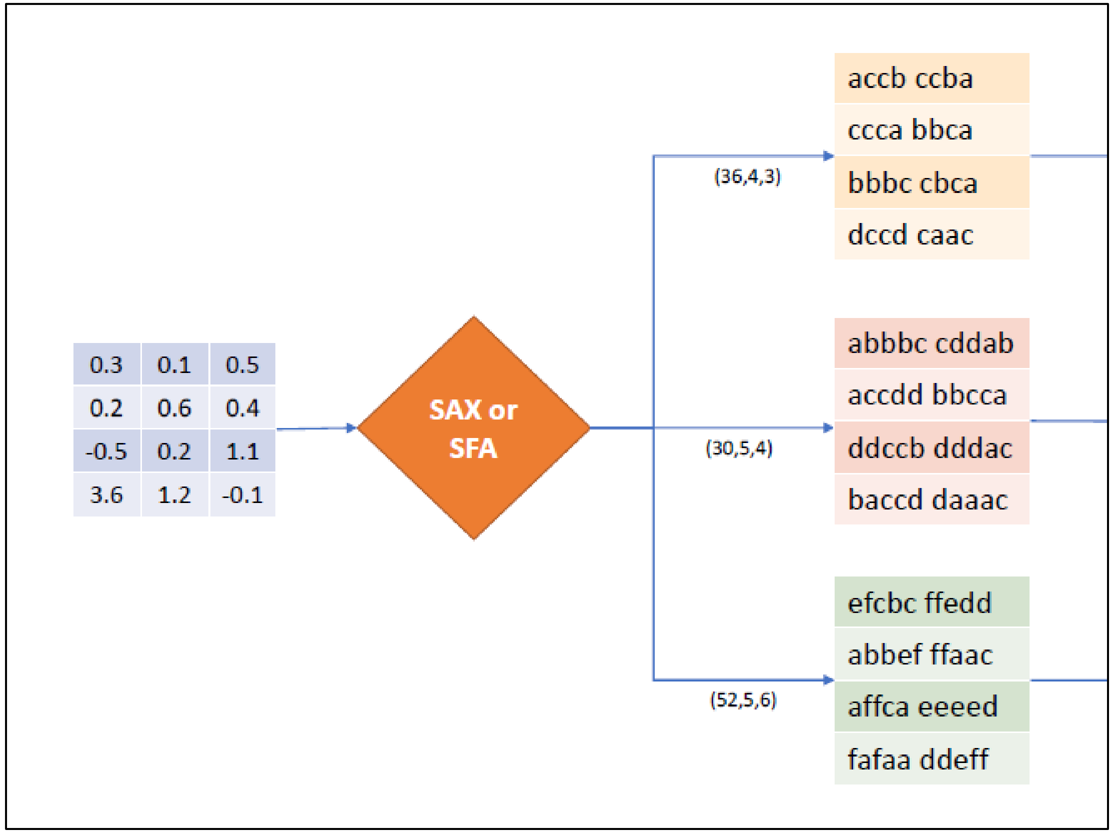
Fig. 1. Workflow for the MrSQM time series classifier with 3 stages: 1. symbolic transform, 2. feature transform, 3. classifier learning.

Methodology

- Symbolic Transformation: SAX or SFA
- Transformation parameters: window size, word length, alphabet size.

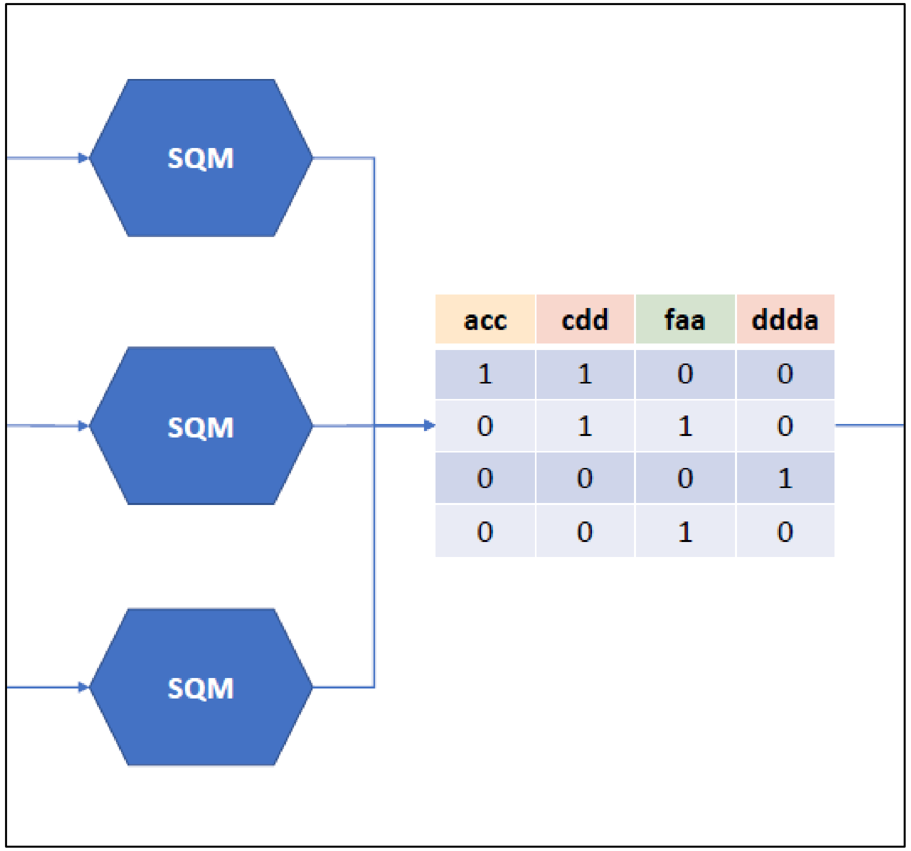
Window size	$2^{3+i/k}$ for I in (0,1,….log(L))
Word length	6,8,10,12,14,16
Alphabet size	3,4,5,6

- Number of symbolic representations: **k*log(L)** where **k** is a hyperparameter and **L** is the length of time series.



Methodology

- SQM: sampling random subsequences.
- Feature values: 1 if the subsequence is found in the time series symbolic representation and 0 if not.
- Optional: Feature selection.
- Concatenating the 0/1 feature vectors from different symbolic representations.



Methodology

Three main building blocks:

- Symbolic transformation: numerical time series to multiple symbolic representations (SAX or SFA).
- Feature transformation: Random subsequences as features.
- Learning algorithm: Logistic regression.

Variants of MrSQM:

- MrSQM-R: Based variants with three stages as above.
- MrSQM-RS: includes an extra step of feature selection after random sampling of features.

Experiments

Experiment Setup:

- Data: **108** fixed-length univariate time series datasets from the UEA/UCR Archive (timeseriesclassification.com)
- System: Linux workstation with an Intel Core i7-7700 Processor and 32GB memory.

Questions:

- Which MrSQM variant is more accurate ? R or RS ?
- Which transformation is more suitable ? SAX or SFA ?
- How is the runtime-accuracy tradeoff ? How does increasing the number of symbolic representations impact the performance ?
- How does MrSQM compare to the state-of-the-art time series classifiers ?

Experiments

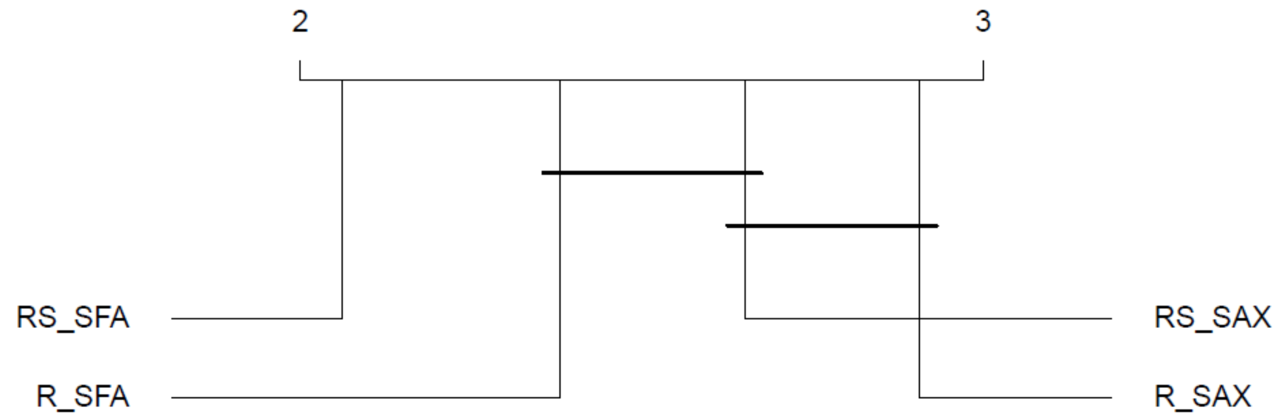
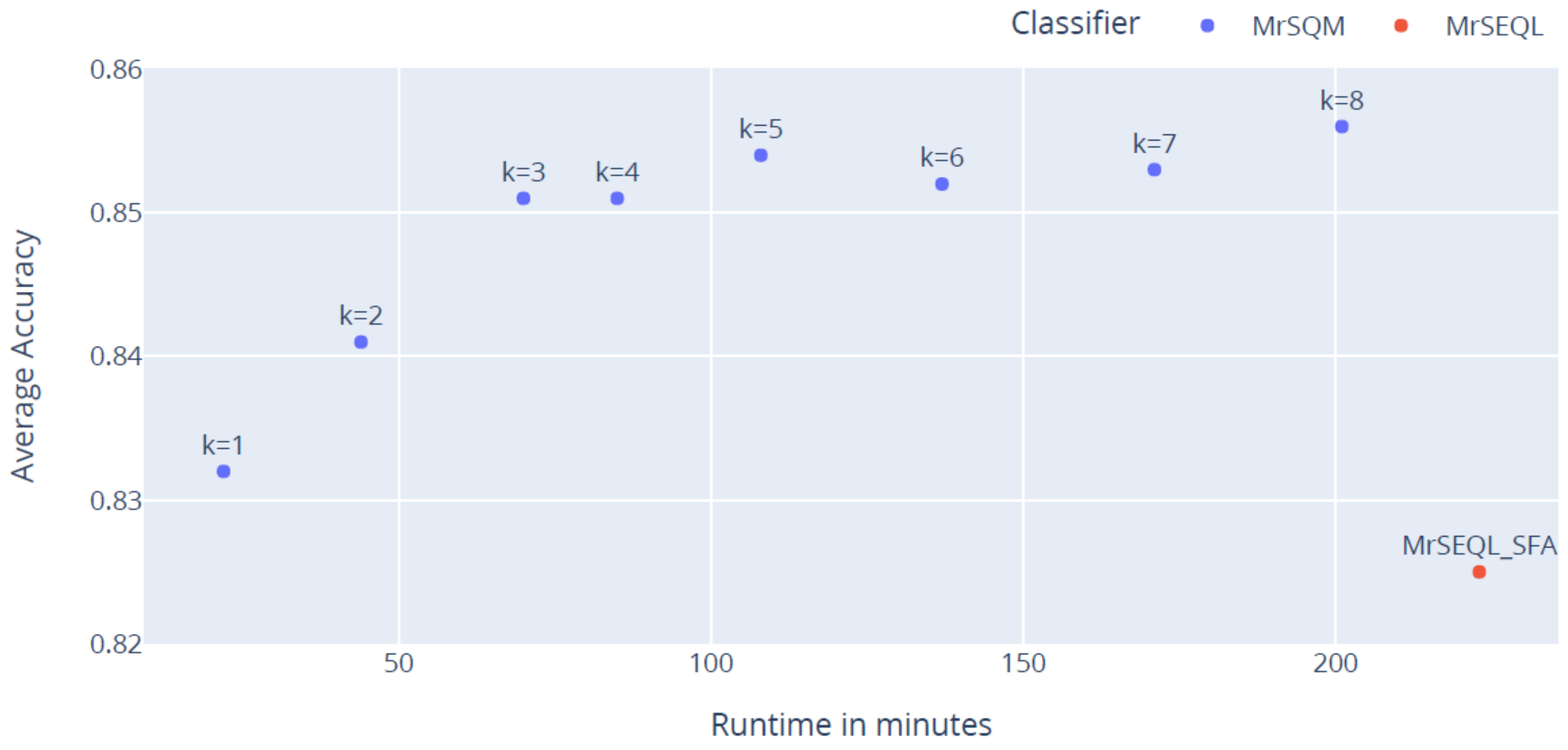


Fig. 3. Comparison of combinations between two variants of MrSQM and two symbolic representations.

Experiments



* MrSQM create **k*log(L)** representations where L is the length of the time series.

Fig. 5. Comparison of average accuracy and total training and prediction time (minutes) for MrSQM-SFA variants at varying k and MrSEQL variants as baseline.

Experiments

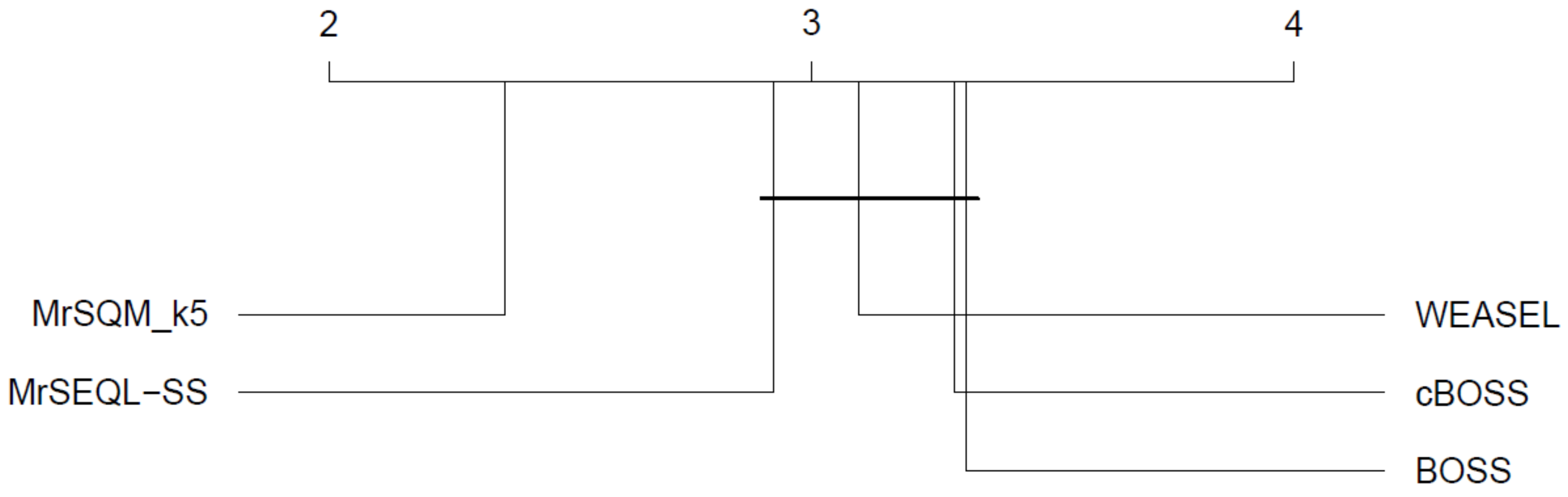


Fig. 7. Comparison of state-of-the-art symbolic time series classifiers across 112 UEA/UCR TSC datasets. The leftmost method has the best average rank.

Experiments

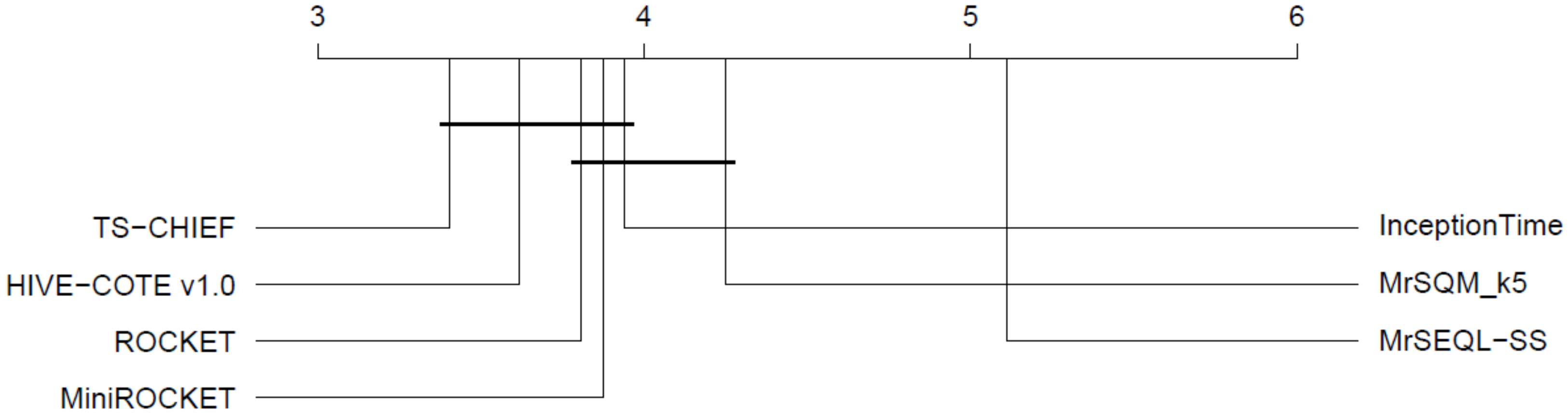
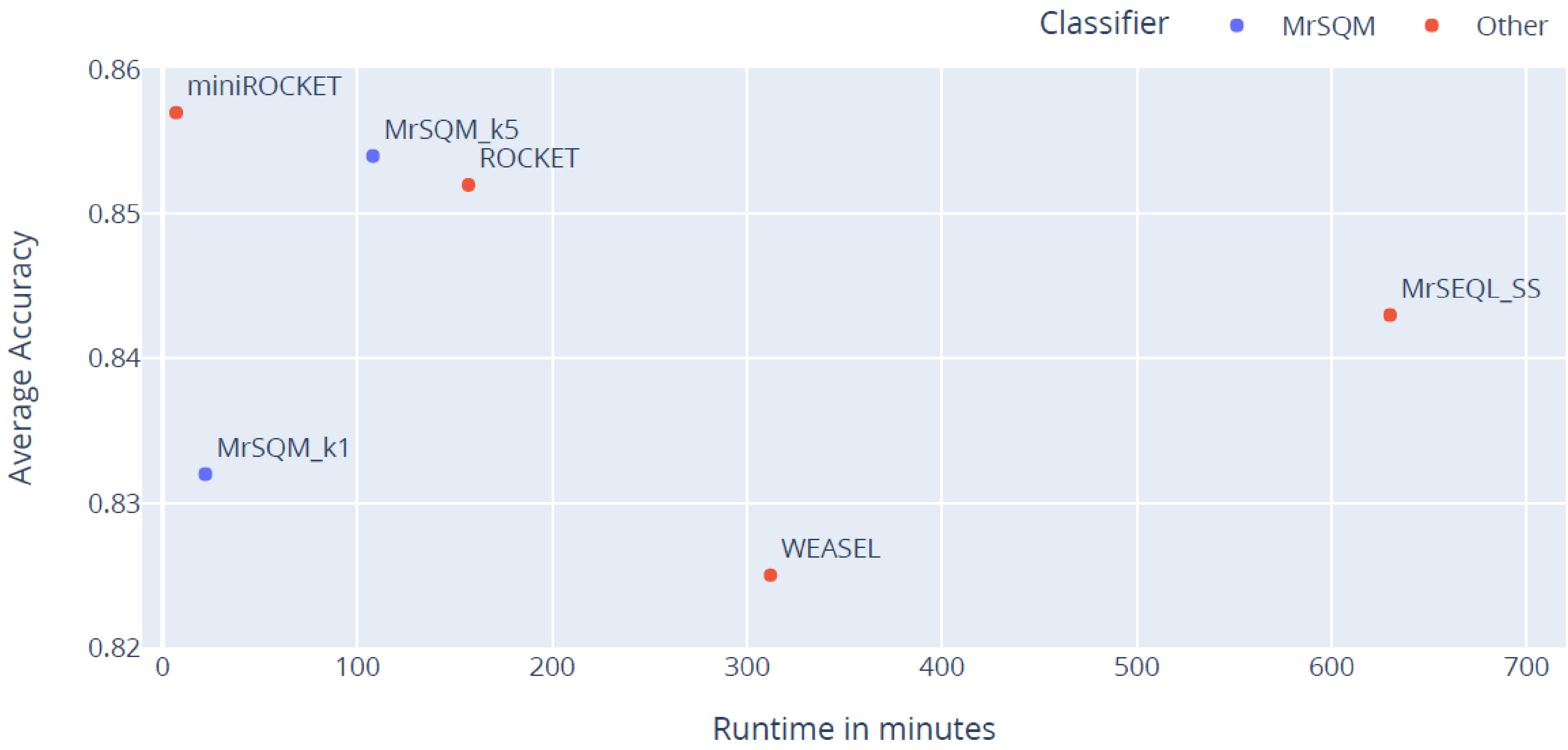


Fig. 8. Comparison with state-of-the-art time series classifiers across 112 UEA/UCR TSC datasets. The leftmost method has the best average rank.

Experiments



Classifier	Total (hours)
MiniROCKET	0.1
MrSQM_K1	0.3
MrSQM_K5	1.5
ROCKET	2.5
WEASEL	5
MrSEQL-SS	10
HIVE-COTE1.0	400
TS-CHIEF	600

Conclusions

- We presented MrSQM, a simple and efficient classifier that achieves state-of-the-art accuracy on the UEA/UCR time series classification benchmark.
- Linear classifiers working in large feature space are very effective.
- Future work: extend MrSQM to multivariate time series.
- All our code and data is publicly available at https://github.com/mlgig/mrsqm

Further Information

Thach Le Nguyen

UCD

Email: thach.lenguyen@ucd.ie

Experiments

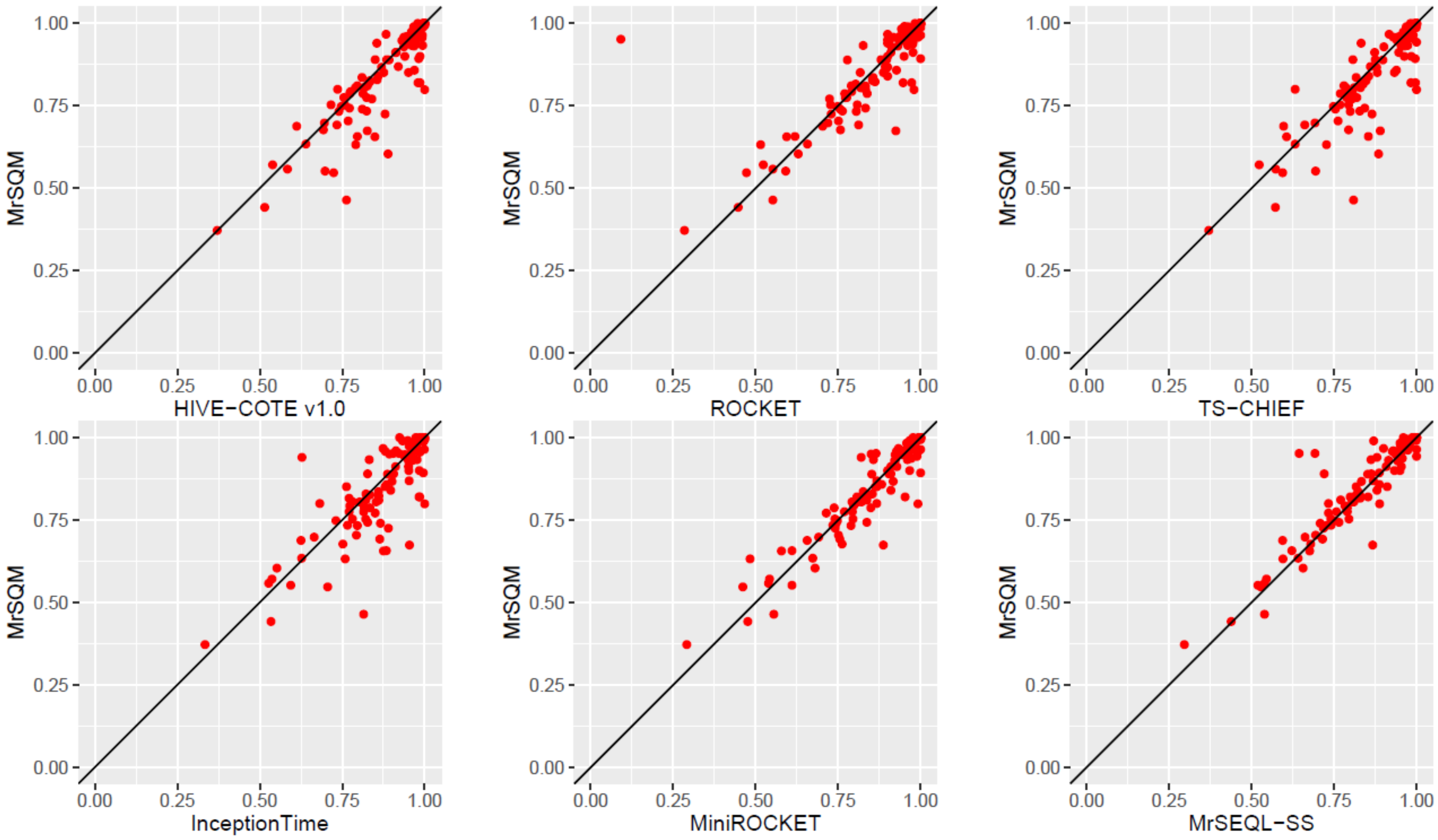


Fig. 9. Pairwise comparison between state-of-the-art time series classifiers and MrSQM with regard to accuracy across 112 UEA/UCR TSC datasets.

Experiments

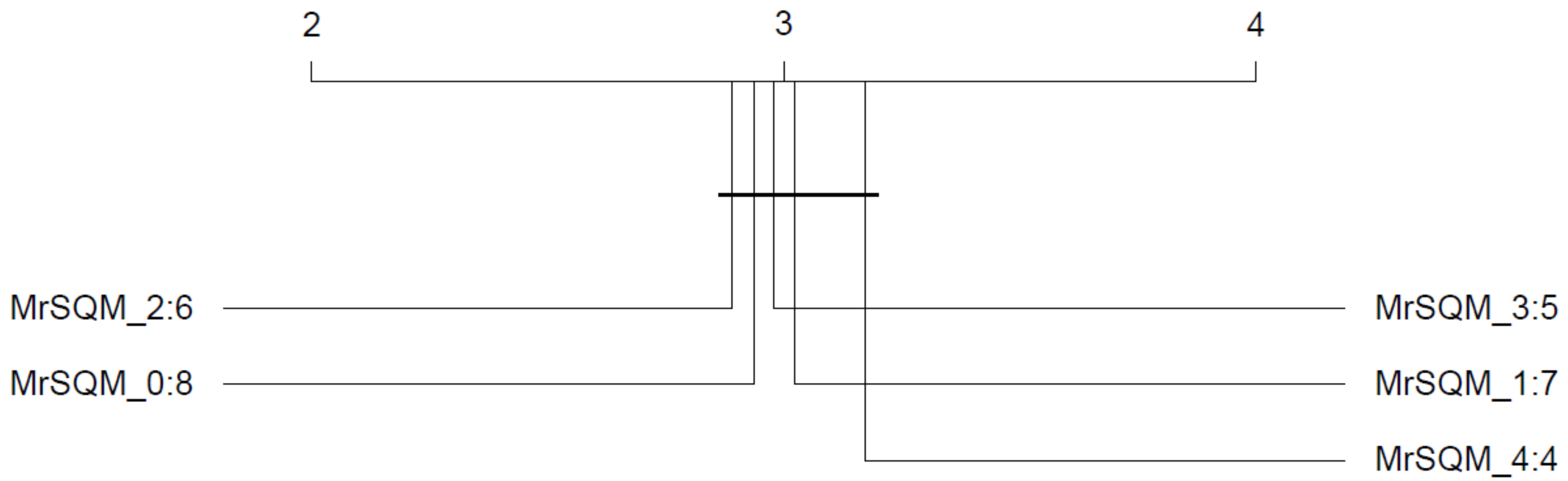


Fig. 6. Comparison between variants of MrSQM with different ratios of SAX and SFA representations. $k_1 : k_2$ means MrSQM generates $k_1 \times log(L)$ SAX representations and $k_2 \times log(L)$ SFA representations.

Experiments