Multivariate Time Series Classification with WEASEL+MUSE

Patrick Schäfer and Ulf Leser

Humboldt University of Berlin, Germany
{patrick.schaefer,leser}@informatik.hu-berlin.de

Abstract. Multivariate time series (MTS) arise when multiple interconnected sensors record data over time. Dealing with this high-dimensional data is challenging for every classifier for at least two reasons: First, an MTS is not only characterized by individual feature values, but also by the interplay of features in different dimensions. Second, the high dimensionality typically adds large amounts of irrelevant data and noise. We present our novel MTS classifier WEASEL+MUSE which addresses both challenges. WEASEL+MUSE builds a multivariate feature vector, first using a sliding-window approach applied to each dimension of the MTS, then extracting discrete features per window and dimension. The feature vector is subsequently fed through feature selection, removing non-discriminative features, and analysed by a machine learning classifier. The novelty of WEASEL+MUSE lies in its specific way of extracting and filtering multivariate features from MTS by encoding context information into each feature. Still, the resulting feature set is small, yet very discriminative and useful for MTS classification. Based on a benchmark of 20 MTS datasets, we found that WEASEL+MUSE is among the most accurate state-of-the-art classifiers.

Keywords: Time series · Multivariate · Classification · Bag-of-Patterns

1 Introduction

A time series (TS) is a collection of values sequentially ordered in time. TS emerge in many scientific and commercial applications, like weather observations, wind energy forecasting, industry automation, mobility tracking, etc. [18] One driving force behind their rising importance is the sharply increasing use of heterogeneous sensors for automatic and high-resolution monitoring in domains such as smart homes [5], machine surveillance [10], or smart grids. A multivariate time series (MTS) arises when multiple interconnected streams of data are recorded over time. These are typically produced by devices with multiple (heterogeneous) sensors like weather observations (humidity, temperature), Earth movement (three axis), or satellite images (in different spectra). We study the problem of multivariate time series classification (MTSC). Given a concrete MTS, the task of MTSC is to determine which of a set of predefined classes this MTS belongs to, e.g., labeling a sign language gesture based on a set of predefined gestures. The high dimensionality introduced by multiple streams of sensors is very challenging for classifiers, as MTS are not only described by individual features but also by their interplay/co-occurrence in different dimensions [1].

In this paper, we introduce our novel domain agnostic MTSC method called *WEASEL+MUSE (WEASEL+MUltivariate Symbolic Extension)*. It conceptually builds on the bag-of-patterns (BOP) [13,12] model and the *WEASEL* [15] pipeline. The BOP model moves a sliding window over an MTS, extracts discrete features per window, and creates a histogram over discrete feature counts. These histograms are subsequently fed into a machine learning classifier. WEASEL+MUSE is different from state-of-the-art classifiers:

1. **Identifiers**: WEASEL+MUSE adds a dimension (sensor) identifier to each extracted discrete feature. Thereby, it can discriminate between the presence of features in different dimensions - i.e., whether the left or right hand was raised.
2. **Derivatives**: To improve accuracy, derivatives in each dimension are added as features to the MTS. These derivatives represent the general shape and are invariant to the exact value at a given time stamp.
3. **Interplay of features**: The interplay of features along the dimensions is learned by assigning weights to features (using logistic regression), thereby boosting or dampening feature counts.
4. **Order invariance**: A main advantage of the BOP model is its invariance to the order of the subsequences, as a result of using histograms over feature counts. Thus, two MTS are similar, if they show a similar number of feature occurrences rather than having the same values at the exact same time instances.
5. **Feature selection**: Given the wide range of features, many non-discriminative features are introduced. We apply statistical feature selection and weighting to identify those features that best discern between classes.

In our experimental evaluation using 20 public benchmark MTS datasets [9] WEASEL+MUSE is constantly among the most accurate methods. WEASEL+MUSE clearly outperforms all other classifiers except for the very recent deep-learning-based method from [6]. The paper is organized as follows: Section 2 briefly recaps definitions. In Section 3 we present related work. In Section 4 we present our novel way of feature generation and selection. Section 5 presents evaluation results and Section 6 our conclusion.

2 Background: Time Series and Bag-of-Patterns

A *univariate* time series (TS) $T = \{t_1, \ldots, t_n\}$ is an ordered sequence of $n \in \mathbb{N}$ real values $t_i \in \mathbb{R}$. A *multivariate* time series (MTS) $T = \{t_1, \ldots, t_m\}$ is an ordered sequence of $m \in \mathbb{N}$ streams (dimensions) with $t_i = (t_{i,1}, \ldots, t_{i,n}) \in \mathbb{R}^n$. For instance, a stream of m interconnected sensors is recording n values at each time instant. As we primarily address MTS generated from automatic sensors with a fixed and synchronized sampling along all dimensions, we can safely ignore time stamps. A time series dataset D contains N time series. Note

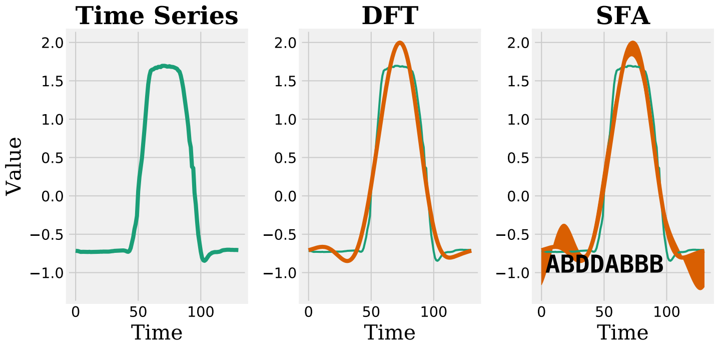

Fig. 1. The Symbolic Fourier Approximation (SFA): A time series (left) is approximated using the truncated Fourier transform (centre) and discretized to the word *ABDDABBB* (right) with the four-letter alphabet ('a' to 'd'). The orange area (right) represents the tolerance for all signals that will be mapped to the same word.

that we consider only MTS with numerical attributes (not categorical). The *derivative* of a stream $t_i = (t_{i,1}, \dots, t_{i,n})$ is given by the sequence of pairwise differences $t'_i = (|t_{i,2} - t_{i,1}|, \dots, |t_{i,n} - t_{i,n-1}|)$. Adding derivatives to an MTS $T = \{t_1, \dots, t_m\}$ of m streams effectively doubles the number of streams: $T = \{t_1, \dots, t_m, t'_1, \dots, t'_m\}$. Given a univariate TS T, a *window* S of length w is a subsequence with w contiguous values starting at offset a in T, i.e., $S(a, w) = (t_a, \dots, t_{a+w-1})$ with $1 \leq a \leq n - w + 1$.

Our method is based on the bag-of-patterns (BOP) model [12,13]. Algorithms following the BOP model build a classification function by (1) extracting subsequences from a TS, (2) discretizing each real valued subsequence into a discrete-valued *word* (a sequence of symbols over a fixed alphabet), (3) building a histogram (feature vector) from word counts, and (4) finally, using a classification model from the machine learning repertoire on these feature vectors. Different discretization functions have been used in literature, including SAX [8] and SFA [14]. SAX is based on the discretization of mean values and SFA is based on the discretization of coefficients of the Fourier transform. Thereby, SFA transforms a real-valued TS window to a word using an alphabet of size c. Figure 1 exemplifies this process for a univariate time series, resulting in the word *ABDDABBB*.

3 Related Work

The techniques used for TSC can broadly be categorized into two classes: (a) similarity-based (distance-based) methods and (b) feature-based methods. *Similarity-based* methods make use of a similarity measure like Dynamic Time Warping (DTW) to compare two TS. In contrast, *feature-based* TSC rely on comparing features, typically generated from substructures of a TS. The most suc-

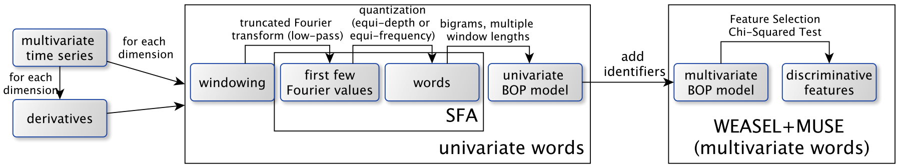

Fig. 2. WEASEL+MUSE Pipeline: Feature extraction, univariate BOP models and WEASEL+MUSE.

cessful approaches are *shapelets* or *bag-of-patterns* (BOP). *Shapelets* are defined as TS subsequences that are maximally representative of a class. For multivariate time series classification (MTSC) some domain agnostic MTSC have been proposed. *Symbolic Representation for Multivariate Time series (SMTS)* [1] uses codebook learning and the bag-of-words (BOW) model for classification. First, a random forest is trained on the raw MTS to partition the MTS into leaf nodes. Each leaf node is labelled by a codebook. For classification a second random forest is trained on the BOW representations. The method *Generalized Random Shapelet Forests (gRSF)* [7] also generates a set of shapelet-based decision trees over randomly extracted shapelets. *Learned Pattern Similarity (LPS)* [2] extracts segments from an MTS, trains regression trees to identify dependencies between segments. It then builds a BOW representation based on the labels of the leaf nodes. Finally, a similarity measure is defined on the BOW representations. *Autoregressive (AR) Kernel* [3] proposes an AR kernel-based distance measure for MTSC. The method *Autoregressive Forests for multivariate time series modelling (mv-ARF)* [16] proposes a tree ensemble trained on autoregressive models, each one with a different lag, of the MTS. *Multivariate LSTM-FCN* [6] introduces a deep learning architecture based on a long short-term memory architecture (LSTM), a fully convolutional network (FCN) and a squeeze and excitation block.

4 WEASEL+MUSE (MUltivariate Symbolic Extension)

WEASEL+MUSE is composed of the building blocks depicted in Figure 2: the symbolic representation SFA [14], BOP models for each dimension, feature selection and the WEASEL+MUSE model. An MTS is first split into its dimensions. Each dimension can then be considered as a univariate TS. To this end, z-normalized windows of varying lengths are extracted from each univariate TS. Next, each window is approximated using the truncated Fourier transform, retaining only the lower frequency components of each window. The extracted Fourier values (real and imaginary part separately) are then discretized into words based on equi-depth or equi-frequency binning using SFA [14]. As a result, words (unigrams) and pairs of words (bigrams) with varying window lengths are computed. These words are concatenated with their identifiers, i.e., the sensor

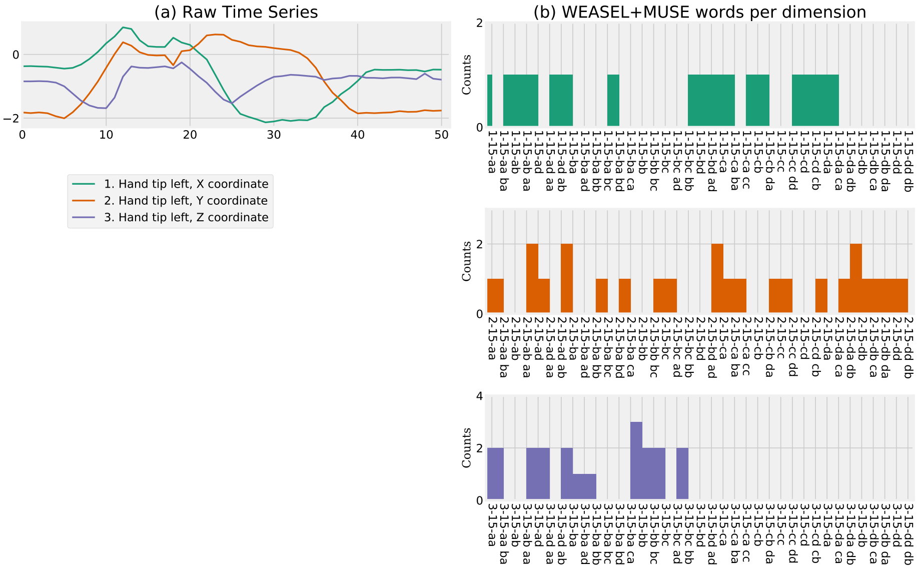

Fig. 3. WEASEL+MUSE model of a motion capture. (a) motion of a left hand in x/y/z coordinates. (b) the WEASEL+MUSE model for each of these coordinates. A feature encodes the dimension, window length and actual word, e.g., 1_15_*aa* for 'left Hand', window length 15 and word 'aa'.

id's dimension and the used window length. Thus, WEASEL+MUSE keeps a disjoint word space for each dimension. Figure 3 exemplifies the WEASEL+MUSE model for a fixed window length 15 on motion capture data. The TS has 3 dimensions (x,y,z coordinates). A feature $('3\_15\_ad', 2)$ represents a unigram 'ad' for the z-dimension with window length 15 and frequency 2.

WEASEL+MUSE supports multivariate time series with streams of variable lengths. When generating features, the window length can be larger than the stream length n. In that case, no features are extracted (equal to feature counts of 0 in the histogram).

Finally, WEASEL+MUSE applies the Chi-squared (χ^2) test to identify the most relevant features. Only features passing a certain threshold are kept to reduce this feature space prior to training the classifier. We set the threshold so that it is high enough for the logistic regression classifier to train a model in reasonable time (and when set too low, training takes longer). We implemented our MTS classifier using liblinear [4] as it scales linearly with the dimensionality of the feature space [11].

The WEASEL+MUSE model is essentially a histogram of discrete features (bag-of-patterns). The logistic regression classifier captures the interplay of features across dimensions by training high weights for characteristic features. Thus, dimensions are not treated separately but the weight vector is trained using features from all dimensions. Still, this approach allows for phase-invariance of

Dataset	SMTS	LPS	mvARF	DTWi	ARKernel	gRSF	MLSTMFCN	MUSE
ArabicDigits	96.4%	97.1%	95.2%	90.8%	98.8%	97.5%	99.0%	**99.2%**
AUSLAN	94.7%	75.4%	93.4%	72.7%	91.8%	95.5%	95.0%	**97%**
CharTrajectories	99.2%	96.5%	92.8%	94.8%	90%	99.4%	99.0%	97.3%
CMUsubject16	99.7%	**100%**	**100%**	93%	**100%**	**100%**	**100%**	**100%**
ECG	81.8%	82%	78.5%	79%	82%	**88%**	87%	**88%**
JapaneseVowels	96.9%	95.1%	95.9%	96.2%	98.4%	80%	**100%**	97.6%
KickvsPunch	82%	90%	97.6%	60%	92.7%	**100%**	90%	**100%**
Libras	90.9%	90.3%	94.5%	88.8%	95.2%	91.1%	**97%**	89.4%
NetFlow	**97.7%**	96.8%	NaN	97.6%	NaN	91.4%	95%	96.1%
UWave	94.1%	**98%**	95.2%	91.6%	90.4%	92.9%	97%	91.6%
Wafer	96.5%	96.2%	93.1%	97.4%	96.8%	99.2%	99%	**99.7%**
WalkvsRun	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
LP1	85.6%	86.2%	82.4%	76%	86%	84%	80%	**94%**
LP2	76%	70.4%	62.6%	70%	63.4%	66.7%	**80%**	73.3%
LP3	76%	72%	77%	56.7%	56.7%	63.3%	73%	**90%**
LP4	89.5%	91%	90.6%	86.7%	96%	86.7%	89%	**96%**
LP5	65%	**69%**	68%	54%	47%	45%	65%	**69%**
PenDigits	91.7%	90.8%	92.3%	92.7%	95.2%	93.2%	**97%**	91.2%
Shapes	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
DigitShapes	**100%**	**100%**	**100%**	**93.8%**	**100%**	**100%**	**100%**	**100%**
Wins/Ties	4	6	4	2	5	6	8	**13**
Mean	90.7%	89.8%	90%	84.6%	88.4%	88.7%	92.1%	**93.5%**
Avg. Rank	4.05	4.05	4.7	6.6	4.35	3.85	2.75	2.45

Table 1. Accuracies for each dataset. The best approaches are highlighted.

features as the classes (events) are represented by the frequency of occurrence of discrete features rather than the exact time instance of an event.

5 Evaluation

Datasets: We evaluated our *WEASEL+MUSE* classifier using 20 publicly available MTS datasets from [9]. Each MTS dataset provides a train and test split which we use unchanged to make our results comparable to prior publications. *Competitors:* We compare WEASEL+MUSE to the 7 domain agnostic state-of-the-art MTSC methods we are aware of: ARKernel [3], LPS [2], mv-ARF [16], SMTS [1], gRSF [7], MLSTM-FCN [6], and the common baseline Dynamic Time Warping independent (DTWi), implemented as the sum of DTW distances in each dimension with a full warping window. All reported numbers in our experiments correspond to the accuracy on the test split. We were not able to reproduce the published results for MLSTM-FCN using their code. The authors told us that this is due to random seeding and their results are based on a single run. Instead, we report the median over 5 runs using their published code [6]. *Training WEASEL+MUSE:* For WEASEL+MUSE we performed 10-fold cross-validation on the train datasets to find the most appropriate parameters for the

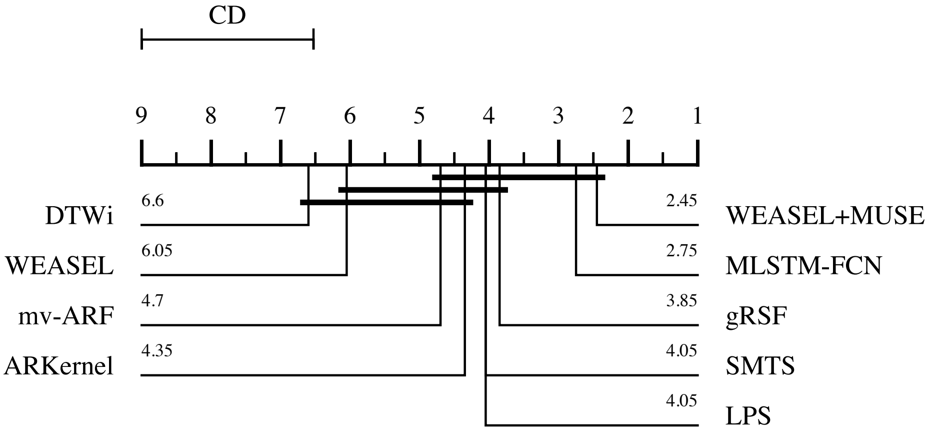

Fig. 4. Average ranks on the 20 MTS datasets. WEASEL+MUSE and MLSTM-FCN are the most accurate.

SFA word lengths $l \in [2, 4, 6]$ and SFA quantization method *equi-depth* or *equi-frequency* binning. We provide the WEASEL+MUSE source code and the raw measurement sheets [17].

5.1 Accuracy

Figure 4 shows a critical difference diagram over the average ranks of the different MTSC methods. Classifiers with the lowest (best) ranks are shown on the right. The group of classifiers that are not significantly different in their rankings are connected by a bar. The critical difference (CD) length at the top represents statistically significant differences. MLSTM-FCN and WEASEL+MUSE show the lowest overall ranks and the highest accuracies. These two are also significantly better than the baseline DTWi. Overall, WEASEL+MUSE has 13 wins (or ties) on the datasets (Table 1), which is the highest of all classifiers. With a mean of 93.5% it also shows the highest average accuracy. Compared to our previous work WEASEL, we see a significant improvement in ranks (2.45 vs. 6.05). WEASEL+MUSE performs best for sensor reading datasets and MLSTM-FCN performs best for motion and speech datasets. Sensor readings are the datasets with the least number of samples N or features n in the range of a few dozens. On the other hand, speech and motion datasets contain the highest number of samples or features in the range of hundreds to thousands. This might indicate that WEASEL+MUSE performs well, even for small-sized datasets, whereas MLSTM-FCN seems to require larger training corpora for the highest accuracy.

6 Conclusion

We have presented a novel multivariate time series classification method following the bag-of-pattern approach and achieving highly competitive classification accuracies. The novelty of WEASEL+MUSE is its feature space engineering using statistical feature selection, derivatives, variable window lengths, bi-grams, and a symbolic representation for generating discriminative words. WEASEL+MUSE offers tolerance to noise (by use of the truncated Fourier transform), phase invariance, and superfluous data/dimensions. In our evaluation on altogether 20 datasets, WEASEL+MUSE is consistently among the most accurate classifiers. It performs well even for small-sized datasets, where deep learning based approaches typically tend to perform poorly.

References

1. Baydogan, M.G., Runger, G.: Learning a symbolic representation for multivariate time series classification. DMKD **29**(2), 400–422 (2015)
2. Baydogan, M.G., Runger, G.: Time series representation and similarity based on local autopatterns. DMKD **30**(2), 476–509 (2016)
3. Cuturi, M., Doucet, A.: Autoregressive kernels for time series. arXiv preprint arXiv:1101.0673 (2011)
4. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. JMLR **9**, 1871–1874 (2008)
5. Jerzak, Z., Ziekow, H.: The DEBS 2014 Grand Challenge. In: Proceedings of the 2014 ACM DEBS. pp. 266–269. ACM (2014)
6. Karim, F., Majumdar, S., Darabi, H., Harford, S.: Multivariate lstm-fcns for time series classification. arXiv preprint arXiv:1801.04503 (2018)
7. Karlsson, I., Papapetrou, P., Boström, H.: Generalized random shapelet forests. DMKD **30**(5), 1053–1085 (2016)
8. Lin, J., Keogh, E.J., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. DMKD **15**(2), 107–144 (2007)
9. Mustafa Gokce Baydogan: Multivariate Time Series Classification Datasets. `http://www.mustafabaydogan.com` (2017)
10. Mutschler, C., Ziekow, H., Jerzak, Z.: The DEBS 2013 grand challenge. In: Proceedings of the 2013 ACM DEBS. pp. 289–294. ACM (2013)
11. Ng, A.Y.: Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In: ICML. p. 78. ACM (2004)
12. Schäfer, P.: Scalable time series classification. DMKD pp. 1–26 (2015)
13. Schäfer, P.: The BOSS is concerned with time series classification in the presence of noise. DMKD **29**(6), 1505–1530 (2015)
14. Schäfer, P., Högqvist, M.: SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In: EDBT. pp. 516–527. ACM (2012)
15. Schäfer, P., Leser, U.: Fast and Accurate Time Series Classification with WEASEL. CIKM pp. 637–646 (2017)
16. Tuncel, K.S., Baydogan, M.G.: Autoregressive forests for multivariate time series modeling. Pattern Recognition **73**, 202–215 (2018)
17. WEASEL+MUSE Classifier Source Code and Raw Results: `https://www2.informatik.hu-berlin.de/~schaefpa/muse/` (2017)
18. Y Chen, E Keogh, B Hu, N Begum, A Bagnall, A Mueen and G Batista : The UCR Time Series Classification Archive. `http://www.cs.ucr.edu/~eamonn/time_series_data` (2015)