Few-shot Time-series Classification with Dual Interpretability

Wensi Tang [*1] **Lu Liu** [*1] **Guodong Long** [*1]

Abstract

Few-shot time-series classification aims to learn a classifier on time series data, and the classifier has a fast-adaptive ability that can categorize unseen samples into a class which receives very few labeled training samples. However, conventional time-series classification algorithms fail to tackle the few-shot scenario. Existing few-shot learning methods are proposed to tackle image or text data, and most of them are neural-based models that lack interpretability. This paper proposes *Dual Prototypical Shapelet Networks (DPSN)*, which not only train a neural network-based model but also interpret the model from the perspective of representative time series samples and shapelets. In particular, the generated dual prototypical shapelets consist of representative samples that can mostly demonstrate the overall shapes of all samples in the class and discriminative partial-length shapelets that can be used to distinguish different classes. We test DPSN on 22 datasets and show that DPSN outperforms state-of-the-art time-series classification methods, especially when trained with few data.

1. Introduction

Training a classification model with very few labeled training samples, namely few-shot classification (Lake et al., 2015). A few-shot classification model should be able to categorize unseen samples to a class according to the generalized concepts and knowledge that is extracted from the very few seen examples (training samples) in this class.

Time-series classification (TSC) has been broadly applied in intelligent-based real-world applications, including heart disease diagnoses (e.g., ECG200 data set), motion detection (e.g., GunPoint data set), traffic analysis (e.g., Melbourne-Pedestrian data set), etc.

However, to the best of our knowledge, seldom research has been done in to solve the TSC problem under few-shot scenario. It can be seen that with the development of domains such as precision medicine, wearable devices, the requirements for few-shot TSC algorithm will increase (Busatto et al., 2008; Sun & Yeh, 2017). An example is building a personalized model for exercise. Although the biomarkers are objective, an individual's feeling is subjective. Which means we need customer's participation to interpenetrate his data. In other word, labeling his data. However, an individual user cannot produce much data, and labeling much data would bring a negative user experience. Thus, few-shot TSC would benefit those areas.

There are two challenges in few-shot TSC. Firstly, conventional state-of-art TSC methods fail to tackle the few-shot scenario. Secondly, existing few-shot learning methods (Chen et al., 2019; Snell et al., 2017; Motiian et al., 2017) are not designed for TSC, and most of them are neural-based models that lack interpretability. Thus, a reasonable interpretation is critically important to convince end-users that the trained few-shot model has a good generalization capability.

This paper proposes a novel Dual Prototypical Shapelet Network (DPSN) that can simultaneously solve the aforementioned few-shot time-series classification problem and address the model interpretability challenge. In particular, the classification model is built on an end-to-end framework that includes a KNN-based lazy learner, and a neural layer which is applied to learn the metrics to transform time series data into a new space. Dual interpretation of the classifier will be generated by two different model interpretability techniques. The first type of interpretation is a selected time-series example, namely a representative shapelet, to be used as a prototype to demonstrate the overall shape of all the samples in a class. The other interpretation output is a sub-sequence of a time-series example, namely discriminative shapelet, that can be used to distinguish different classes. The framework of DPSN is shown in Figure 1.

The papers main contributions are summarised as follows:

- The first study solves the few-shot time-series classification task with neural-based methods.
- DPSN could reveal the overall shape of all the samples in a class
- DPSN could reveal discriminative shapelet between classes

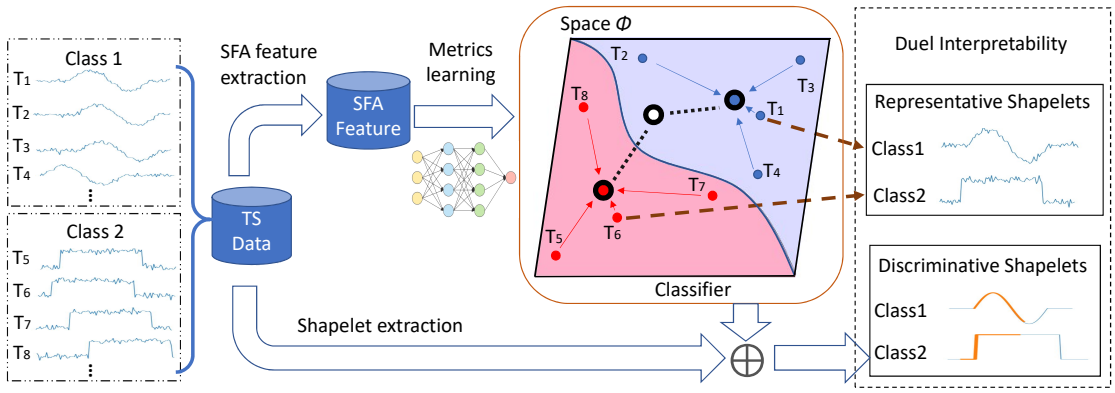

Figure 1. The framework of DPSN. The framework contains three components: 1) the feature extraction will transform Time series into SFA feature and Shapelet feature; 2) The classification part will build a prototype of each class by aggregating SFA feature via a metric learning neural network; 3) The dual interpretability has two-fold explanations including representative shapelet which was identified by prototype from classification part, and discriminative shapelets which can be learned by combining information from shapelets feature and prototype.

2. Related Works

2.1. Time series classification

From the University of California, Riverside (UCR) (Dau et al., 2018) time series classification archive, COTE (Lines et al., 2016), WEASEL (Schäfer & Leser, 2017), ST (Hills et al., 2014) and BOSS (Schäfer, 2015) are top four state-of-the-art algorithms for TSC. Due to the reason that COTE is an ensemble learning method, we do not consider it. Thus, We choose the other three as our baseline to compare accuracy and precision.

Besides accuracy, ST is also remarkable for its interpretability. Thus, we compared our interpretability with ST.

2.2. Few-shot Learning

Few-shot learning aims at training a model with very few labeled training samples. In computer vision and nature language processing domains, many solutions has been introduced to overcome the few-shot problem. (Chen et al., 2019; Finn et al., 2017) aims to learn the fast adaptation ability to new tasks with few training samples. (ZHANG et al., 2018; Dixit et al., 2017) use generative models to create more training samples to offset the few-shot problem.

However, those existing few-shot learning algorithms are designed for image or text classification related tasks. Thus we need to mitigate them to TSC.

2.3. Interpretability for Neural-based Model

Interpretability is critically important for neural-based models due to its black-box nature. Conventional model (Fisher et al., 2018; Kim et al., 2016; Koh & Liang, 2017) interpretation analyses the models parameters and outcomes to generate an interpretation for the model.

Gradients analysis is also important for model interpretation, e.g. Layer-Wise Relevance Propagation (Bach et al., 2015), and gradient sensitivity analysis (Montavon et al., 2018). Many research work has tried to modify the existing DNN/CNN/RNN-based or attention-based neural model to generate intermediate information for model interpretation.

3. Dual Prototypical Shapelet Networks

Following most recent time series learning works (Zhang et al., 2016; Schäfer & Leser, 2017; Bagnall et al., 2017), we learn from K classes training time series $\mathcal{T} = \{T_1, ..., T_N\}$ annotated by labels $\mathcal{Y} = \{y_1, ..., y_N\}$. We undertake classification in more challenging scenarios where the number of training samples for every class is reduced to 2, i.e., $\|\mathcal{T}^k\|$=2. We do not test when $\|\mathcal{T}^k\|$=1 for it makes no sense to learn a prototype from one sample without prior knowledge. The overall framework of the proposed DPSN method is shown in Figure 1. The whole procedure is divided into two stages: feature extraction using SFA (Sec. 3.1) and classification using a prototypical network (Sec. 3.2).

3.1. SFA Feature Extraction

We followed this paper (Schäfer, 2015) to transform the time series to dense SFA word histogram features i.e.,

$\mathcal{T} \rightarrow \mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$, for classification.

3.2. Prototypical Network

We use a prototypical network to solve the few-shot time series classification problem. The prototype network is a

KNN based classifier, which classifies samples to the class with the nearest prototype. Prototype $\boldsymbol{C}_k$ for class k is defined by the feature average of $\mathcal{X}^k$:

$$\boldsymbol{C}_k = \frac{1}{\|\mathcal{X}^k\|} \sum_{x \in \mathcal{X}^k} f_\phi(\boldsymbol{x}), \tag{1}$$

where $\mathcal{X}^k$ denotes SFA features of all training samples in class k, $\boldsymbol{x}$ is SFA feature for a time series and $f(\cdot)$ is the feature extractor parameterised by ϕ.

Classification result is calculated as a softmax over all classes $k' \in \mathcal{Y}$ using euclidean distance between a sample and every prototypes:

$$\Pr(y = k|\boldsymbol{x}) = \frac{\exp(-\|f_\phi(\boldsymbol{x}) - \boldsymbol{C}_k\|^2)}{\sum_{k'} \exp(-\|f_\phi(\boldsymbol{x}) - \boldsymbol{C}_{k'}\|^2)} \tag{2}$$

.

The cross entropy loss$L$ of data sampled from the distribution $\mathcal{D}$ can be formulated as:

$$L = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} - \log \Pr(y|\boldsymbol{x};\phi) \tag{3}$$

3.3. Interpret model using representative shapelets

We choose a representative example by finding the time series example which is the nearest neighbor of the prototype in space ϕ. This representative sample is the prototype approximation that can demonstrate the overall shapes of all supportive samples in the class. These representative shapelets for classes are a type of explanation for the trained DPSN classifier.

Given a class k, the prototype is generated by simply calculating the weighted sum of all supportive samples in the class, and the representative sample is selected using the following equation

$$T_i, i = \arg\min_i F(f_\phi(x_i), \boldsymbol{C}_k) \tag{4}$$

where F is the distance function which could be Euclidean distance or another, e.g. cosine dissimilarity, or KL divergence. The distance calculation is based on the representation in space ϕ. Thus, in our work, we use Euclidean distance

3.4. Interpret model using discriminative shapelets

Given a class, its discriminative shapelet should be 1) exists in every same-class sample, and 2) does not exist in any diff-class sample. We denote the representative sample by T and extract many sub-sequence T^i using a slider window. The shapelet of T^i is S^i, discriminative shapelets are denoted as S.

Given the class k, its best or optimal discriminative shapelet S_k can be found by maximising the likelihood in Equation 5.

$$S_k = \arg\max_i L_o(S^i, \mathcal{T}, \mathcal{Y}^k) \tag{5}$$

where $\mathcal{T}$ and $\mathcal{Y}^k$ combine to product a bi-classification dataset where the positive class is the given class k and the negative class is all other classes. The likelihood function needs to be aligned to the aforementioned two requirements of discriminative shapelets. In particular, we measure the likelihood using f-test as demonstrated in Equation 6.

$$L_o = \frac{between\ class\ variability}{within\ class\ variability} \tag{6}$$

where the large value of "between-class variability" indicates that the selected shapelet does "not exist" in the diff-class samples, and the small value of the "within-class variability" indicates that the shapelet is likely to "exist" in the same-class samples. The shapelet has the largest f-test value is the discriminative shapelet.

The "between-class variability" measures the distance between different classes using the following equation.

$$\frac{\sum_{i=1}^{K} N_i * (\overline{d_i} - \overline{d})^2}{K - 1} \tag{7}$$

where N_i is the number of samples i the $i-th$ class, $\overline{d_i}$ denotes the mean value of all samples in the $i-th$ class, $\overline{d}$ denotes the overall mean of all classes, and K is the number of classes that is 2 in the bi-classification scenario.

The "within-class variability" calculates the variance of all the distance-values in the same class using the following equation.

$$\frac{\sum_{i=1}^{K} \sum_{j=1}^{N_i} (d_{ij} - \overline{d_i})^2}{N - K} \tag{8}$$

As discussed, the d is the value to estimate the likelihood. d is the distance between a shapelet and a time-series example. In particular, a time-series T will be transformed to a set of subsequences T^i that is the same length as shapelet S. As we expect to check whether or not the shapelet exists in the time-series, distance d_{ST} is the minimal distance between S and the set of subsequences T^i. as calculated by Equation 9.

$$d_{ST} = \min_{i=1..ns} F(T^i, S) \tag{9}$$

where F is the distance function, ns is the number of generated sub-sequences from time-series T.

4. Experiments

4.1. Experiment Setup

To, simulate few-shot scenario, we select datasets with few training samples. To easily visualize and interpret the time-series, we select datasets with short time series. To fulfill

these two requirements, we choose the top 1/6 of the 128 datasets with the shortest total length of all time series samples (the number of training sample * the length of the dataset). In total, we choose 22 datasets.

Those datasets have training dataset and test dataset. We re-sample from training dataset to simulate the few-shot scenario. we set the sample ratio number from 0.1 to 0.9 with the step of 0.1. Thus, for one UCR dataset, we can build 9×10 few-shot datasets from it. We will do an experiment on each few-shot dataset.

Our method uses SFA feature and an optimized shapelets method. Thus, we compare with the state-of-art methods which uses SFA feature (WEASEL BOSS) or shapelets (ST).

The transformation network has two fully connected layers(hidden layer to 256 and the output dimension to 64) The the SFA feature was same setting with (Schäfer, 2015) The setting for prototypical neural network is follows: Adam to train our model with the learning rate of 0.002, the learning rate decay of 1e-5, the epochs number of 1000, and the momentum of 0.7. We anonymously public our code here[1].

4.2. Experiment Result

4.2.1. Experiment Result of Classification

Due to the large size of results, we can only list the average result in this paper, and the full result is here [2]. In Figure 2, the average accuracy shows that when we only have few samples, DPSN could outperform other baselines. From the average accuracy STD perspective, our method is more robust than other baselines.

It should be noticed that when the sample ratio is 0.1 the STD of WEASEL are better than us and the average accuracy are not of much difference. This is because at this sample ratio, the WEASEL will not work in some dataset for the algorithm needs at least two training samples to run. Thus, the result of WEASEL is not the actual result.

4.2.2. Interpretability of DPSN

Due to the page limitation, we could only show the interpretability result of one dataset. The other results are here [3]. We choose the BME dataset for it a handcraft dataset which is easy for people to understand without background knowledge. For representative shapelets, as it is shown in Figure 3, we can see that most time series in class 0 and class 2 have smaller values in the middle. This kind of pattern is similar to our representative shapelets. In other words, the data we choose is more representative of its class.

For discriminative shapelets, the ST does not consider the location of shapelet. Thus, it cannot find the high spike in the third image as DPSN. From the interpretability result, we could know that the difference between Class 0 and Class 2 is at the end of the time series. The character of class 1 is it's high value in the middle and not spike (which can be seen by the representative shapelets).

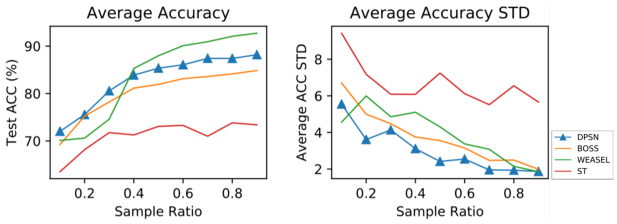

Figure 2. The average accuracy and variance over 22 datasets. Our algorithm outperforms three baselines with lower variance when using few data (Sample ratio ≤ 0.4).

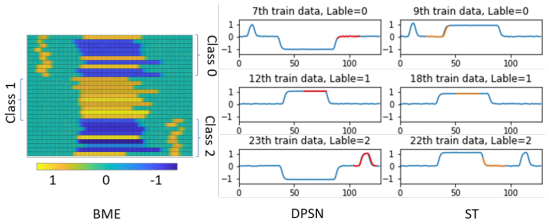

Figure 3. The left image is the visualization of BME dataset. Images in the middle are the DPSN result. The blue line is the representative training sample for each class. The red line is the discriminative shapelet. The left images are the result of ST. The orange line is shapelet picked by ST. The blue line is which sample those shapelets were learned.

5. Conclusions

We propose a novel Dual Prototypical Shapelet Networks for few-shot time-series classification. To the best of our knowledge, we are the first to solve time series classification under the scenario of few-shot using neural network-based model. In experiments, DPSN ourperforms three state-of-the-art baselines in 22 datasets especially when the amount of data significantly decreases. Moreover, we also interpret DPSN from two perspectives: representative sample and discriminative shapelets.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for

[1] https://github.com/IJCAI2019-985/DPSN

[2] https://github.com/IJCAI2019-985/DPSN/blob/master/results/result.txt

[3] https://github.com/IJCAI2019-985/DPSN/blob/master/SFA_Python-master/test/image_result/

non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.

Busatto, G. F., Diniz, B. S., and Zanetti, M. V. Voxel-based morphometry in alzheimers disease. *Expert review of neurotherapeutics*, 8(11):1691–1702, 2008.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification. In *ICLR*, 2019.

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

Dixit, M., Kwitt, R., Niethammer, M., and Vasconcelos, N. Aga: Attribute-guided augmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Fisher, A., Rudin, C., and Dominici, F. Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective. *arXiv preprint arXiv:1801.01489*, 2018.

Hills, J., Lines, J., Baranauskas, E., Mapp, J., and Bagnall, A. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4):851–881, 2014.

Kim, B., Khanna, R., and Koyejo, O. O. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pp. 2280–2288, 2016.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1885–1894. JMLR. org, 2017.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Lines, J., Taylor, S., and Bagnall, A. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1041–1046. IEEE, 2016.

Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Motiian, S., Jones, Q., Iranmanesh, S., and Doretto, G. Few-shot adversarial domain adaptation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6670–6680. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7244-few-shot-adversarial-domain-adaptation.pdf.

Schäfer, P. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.

Schäfer, P. and Leser, U. Fast and accurate time series classification with weasel. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 637–646. ACM, 2017.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Sun, J. C.-Y. and Yeh, K. P.-C. The effects of attention monitoring with eeg biofeedback on university students' attention and self-efficacy: The case of anti-phishing instructional materials. *Computers & Education*, 106: 73–82, 2017.

Zhang, Q., Wu, J., Yang, H., Tian, Y., and Zhang, C. Unsupervised feature learning from time series. In *IJCAI*, pp. 2322–2328, 2016.

ZHANG, R., Che, T., Ghahramani, Z., Bengio, Y., and Song, Y. Metagan: An adversarial approach to few-shot learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2365–2374. Curran Associates, Inc., 2018.