Scale-teaching: Robust Multi-scale Training for Time Series Classification with Noisy Labels

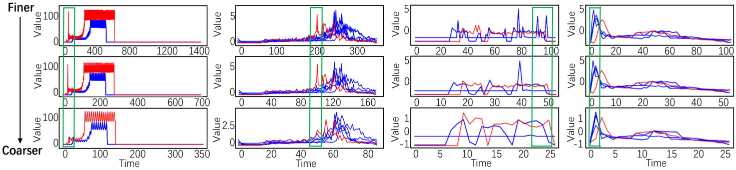
Zhen Liu[1], Peitian Ma[1], Dongliang Chen[1], Wenbin Pei[2], Qianli Ma[1]

[1]South China University of Technology, Guangzhou, China
[2]Dalian University of Technology, Dalian, China

Outline

Motivation

- Time series classification has recently received much attention in deep learning. To improve the robustness of DNNs against noisy labels, existing methods for image data regard samples with small training losses as correctly labeled data.
- However, the discriminative patterns of time series are easily distorted by external noises during the recording process. For example, in a smart grid, distortions may occur due to sampling frequency perturbations, imprecise sensors, or random differences in energy consumption.



(a) PLAID (b) Lightning7 (c) ElectricDevices (d) MedicalImages

Figure 1: Illustration of time series samples *from the same category* at different time scales. Among all samples in the same category, red indicates the one with the largest variance, and blue indicates a few samples with the smallest variance.

Contribution

- We propose a **cross-scale fusion mechanism** to help the model select more reliable clean labels by exploiting complementary information from different scales (Figure 2 (a)).
- We further introduce **multi-scale embedding graph learning** for noisy label correction, using well-learned multi-scale time series embeddings at sample feature levels (Figure 2 (b)).
- Extensive experiments on multiple benchmark time series datasets show that the proposed Scale-teaching paradigm achieves a state-of-the-art classification performance and robustness.

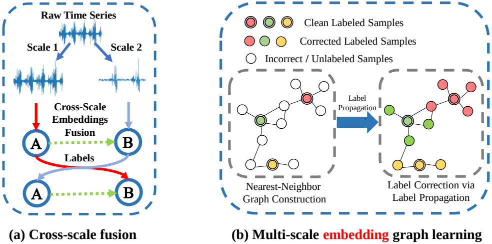


Figure 2: The core contributions of the proposed Scale-teaching paradigm.

Scale-teaching: Model Architecture

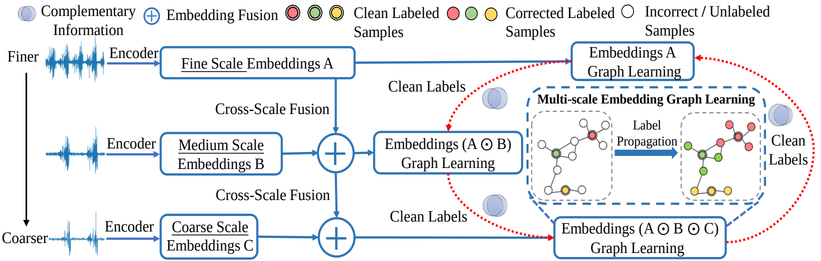


Figure 3: The Scale-teaching paradigm's general architecture comprises two core processes: (i) clean label selection and (ii) noisy label correction.

Experimental Setup

- We conduct experiments utilizing three time series benchmarks (four individual large datasets, UCR 128 archive, and UEA 30 archive). The UCR archive contains 128 univariate time series datasets from different real-world scenarios. The UEA archive contains 30 multivariate time series datasets from real-world scenarios.
- We use three types of noisy labels on the training set for evaluations, namely Symmetric noise, Asymmetric noise, and Instance-dependent noise. Like existing work, we use the test set with correct labels for evaluations.
- We compare Scale-teaching with seven approaches: 1) **Standard**, 2) **Mixup**, 3) **Co-teaching**, 4) **FINE**, 5) **SREA**, 6) **SELC**, and 7) **CULCU**. Among them, Standard, Mixup, and Co-teaching are the benchmark methods for label-noise learning. FINE, SELC, and CULCU are the state-of-the-art methods that do not need to focus on data types, and SREA is the state-of-the-art method in time series domain.

Main Results

Table 1: Test classification accuracy results compared with baselines on three time series benchmarks. The best results are **bold**, and the second best results are underlined.

Dataset	Noise Ratio	Metric	Standard	Mixup	Co-teaching	FINE	SREA	SELC	CULCU	Scale-teaching
Four individual large datasets	Sym 20%	Avg Rank	4.75	4.75	4.50	7.50	6.50	4.50	2.50	**1.00**
	Sym 50%	Avg Rank	4.75	4.50	4.75	7.25	5.75	4.50	3.25	**1.25**
	Asym 40%	Avg Rank	5.00	5.50	3.75	7.50	5.75	4.00	3.25	**1.00**
	Ins 40%	Avg Rank	4.75	4.25	4.25	7.25	6.00	4.75	3.50	**1.00**
UCR 128 archive	Sym 20%	Avg Rank	4.15	4.33	3.61	7.50	6.16	3.48	3.54	**3.02**
		P-value	1.90E-04	4.06E-05	1.90E-03	1.49E-34	1.70E-17	3.04E-03	8.57E-03	-
	Sym 50%	Avg Rank	4.31	4.57	4.05	6.43	5.89	3.56	3.86	**3.11**
		P-value	3.15E-05	1.70E-05	4.02E-04	7.48E-19	1.22E-15	1.40E-02	4.93E-03	-
	Asym 40%	Avg Rank	4.38	4.80	3.93	6.91	5.91	3.30	3.67	**2.95**
		P-value	1.62E-05	3.53E-07	6.10E-04	1.93E-23	9.82E-14	1.89E-02	2.24E-02	-
	Ins 40%	Avg Rank	4.05	4.52	4.02	7.04	6.18	3.30	3.77	**2.95**
		P-value	1.43E-05	1.81E-06	2.43E-04	9.81E-26	2.36E-17	3.27E-02	1.54E-02	-
UEA 30 archive	Sym 20%	Avg Rank	5.03	5.20	3.83	6.37	4.77	3.73	4.00	**2.73**
		P-value	6.61E-04	3.33E-04	2.69E-02	2.37E-05	1.14E-02	2.63E-02	3.93E-02	-
	Sym 50%	Avg Rank	5.17	5.73	4.23	6.23	3.93	3.83	4.30	**2.43**
		P-value	2.98E-04	7.40E-05	1.59E-02	9.35E-05	1.67E-02	1.08E-02	3.75E-02	-
	Asym 40%	Avg Rank	5.60	4.77	4.40	6.13	4.20	4.00	3.97	**2.73**
		P-value	3.81E-03	6.17E-03	1.63E-02	9.33E-05	1.36E-02	2.62E-02	3.88E-02	-
	Ins 40%	Avg Rank	5.20	4.77	4.33	6.60	4.27	4.20	3.77	**2.60**
		P-value	6.08E-04	2.92E-03	1.20E-02	2.55E-05	5.52E-03	1.08E-02	3.47E-02	-

Multi-scale Analysis

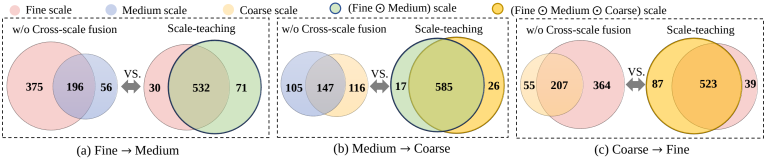


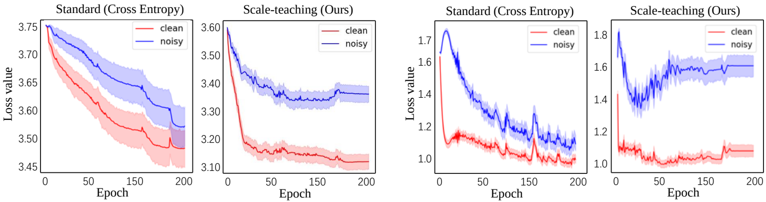
Figure 4: Venn diagram of the average number of correctly classified samples for the different scale sequences of UCR 128 archive with Sym 20% noisy labels. The numbers in the figure indicate the complements and intersections of classification results at different scales.

Multi-scale Analysis

Table 2: The test classification accuracy (%) results of different scale classifiers on UCR 128 archive. The best results are **bold**, and the second best results are underlined.

Method		w/o Cross-scale fusion			Scale-teaching		
Noise Ratio	Metric	Fine	Medium	Coarse	Fine	Medium	Coarse
Sym 20%	Avg Acc	65.13	30.11	28.17	59.67	<u>68.17</u>	**68.70**
	Avg Rank	2.38	5.09	5.37	3.20	<u>2.17</u>	**2.11**
	P-value	1.89E-03	2.85E-37	2.07E-40	1.58E-09	3.74E-02	-
Asym 40%	Avg Acc	49.61	29.01	28.87	47.75	<u>51.93</u>	**52.87**
	Avg Rank	2.64	4.78	4.75	3.01	<u>2.45</u>	**2.27**
	P-value	1.94E-03	6.78E-25	1.59E-27	1.80E-07	2.80E-02	-

Small-loss Analysis



(a) NonInvasiveFetalECGThorax1 (b) ECG5000

Figure 5: The change of loss values for clean and noisy time series samples under Aysm 40% noise labels. The solid line and shading indicate the mean and standard deviation loss values of all clean (or noisy) training samples within each epoch.

Ablation Study

Table 3: The test classification accuracy (%) results of ablation study (values in parentheses denote drop accuracy).

Method	HAR		UniMiB-SHAR	
	Sym 50%	Asym 40%	Sym 50%	Asym 40%
Scale-teaching	**90.17**	**89.62**	**81.31**	**70.68**
w/o cross-scale fusion	88.47 (-1.70)	87.64 (-1.98)	73.32 (-7.99)	61.62 (-9.06)
only single scale	89.01 (-1.06)	88.11 (-1.51)	69.89 (-11.42)	60.32 (-10.36)
w/o graph learning	88.06 (-2.11)	87.65 (-1.97)	79.72 (-1.59)	68.87 (-1.81)
w/o moment	89.76 (-0.41)	88.76 (-0.86)	80.57 (-0.74)	69.85 (-0.83)
w/o dynamic threshold	89.12 (-1.05)	88.75 (-0.87)	77.42 (-3.89)	69.53 (-1.15)

Conclusion

- **Limitations**: The input scales of our proposed Scale-teaching paradigm can only select a fixed number of scales for training, and the running time will increase as the number of scales increases.
- This paper proposes a deep learning paradigm for time-series classification with noisy labels called Scale-teaching. Specifically, we propose cross-scale fusion and multi-scale graph learning for selecting clean labels and noisy label correction, respectively.
- Experiments on the three time series benchmarks show that the Scale-teaching paradigm can utilize the multi-scale properties of time series to effectively handle noisy labels.
- In the future, we will explore the design of scale-adaptive time-series label-noise learning models.