Communication-Efficient Ridge Regression in Federated Echo State Networks

Valerio De Caro, Antonio Di Mauro, Davide Bacciu and Claudio Gallicchio *

University of Pisa - Department of Computer Science
Largo Bruno Pontecorvo, 3, 56127, Pisa - Italy

Abstract. Federated Echo State Networks represent an efficient methodology for learning in pervasive environments with private temporal data due to the low computational cost required by the learning phase. In this paper, we propose Partial Federated Ridge Regression (`pFedRR`), an approximate, communication-efficient version of the *exact* method for learning the readout in a federated setting. Each client compresses the local statistics to be exchanged with the server via an importance-based method, which selects the most relevant neurons with respect to the local distribution. We evaluate the methodology on two Human State Monitoring benchmarks, and results show that the importance-based selection of the information significantly reduces the communication cost, while acting as a regularization method to improve the generalization capabilities.

1 Introduction

In pervasive computing environments, human-centric cyber-physical systems generate vast amounts of temporal data, and are characterized by issues such as privacy, security, and communication bandwidth limitations. This has led to the emergence of federated learning, a decentralized approach where models are trained on data distributed across multiple devices without the need to transfer the data to a central server. In such setting, Federated Echo State Networks (ESNs) were proven to be effective thanks to their capability to efficiently handle temporal data, as well as for the low computational cost of the learning algorithms [1, 2]. In this work, we aim to further improve `FedRR` [1], an algorithm that performs an exact computation of the global readout in a federated setting, towards communication efficiency.

To do so, we propose Partial Federated Ridge Regression (`pFedRR`), where each client applies a policy to select a subset of parameters to be exchanged with the server with negligible computational overhead. Additionally, we propose an importance-based policy that selects the neurons which are most relevant with respect to the local distribution. We evaluate the algorithm with the proposed policy on two Human State Monitoring benchmarks by comparing it with two baselines: (1) `FedRR`; (2) `pFedRR` with a random policy. We show that, depending on the availability of clients and the distribution of the data among clients, selecting the most important units leads to better generalization capabilities while reducing communication cost.

*This work is supported by the TEACHING project funded by the EU Horizon 2020 under GA n. 871385.

2 Compressed Federated Ridge Regression

Echo State Networks. Reservoir Computing (RC) [3] paradigm avoids the shortcomings of gradient-descent in Recurrent Neural Networks (RNNs) training by using two main components: a *reservoir*, a recurrent layer of sparsely connected neurons, holding the internal state which evolves over time; a *readout*, a linear transformation on the domain of the reservoir states. Echo State Networks [4] represents one of the pioneering reservoir computing methods. The approach is based on the assumption that if the recurrent layer is characterized by *stable dynamics*, training only a linear readout from it is often sufficient to achieve excellent performance in practical applications. Formally, given an input sequence $\mathbf{u}(t) \in \mathbb{R}^{N_u}$ for $t \in \{0, 1, \dots, T\}$, the state transition function of the discrete-time dynamical system and the readout can be described as

$$\begin{aligned} \mathbf{x}(t) &= (1-a)\,\mathbf{x}(t-1) + af\left(\mathbf{W}_{in}\mathbf{u}(t) + \mathbf{b}_{rec} + \hat{\mathbf{W}}\mathbf{x}(t-1)\right) \\ \mathbf{y}(t) &= \mathbf{W}\mathbf{x}(t) + \mathbf{b}_{out} \end{aligned} \tag{1}$$

where $\mathbf{W}_{in} \in \mathbb{R}^{N_x \times N_u}$ is the input-to-reservoir weight matrix, $\hat{\mathbf{W}} \in \mathbb{R}^{N_x \times N_x}$ is the recurrent reservoir-to-reservoir weight matrix, $\mathbf{W} \in \mathbb{R}^{N_y \times N_x}$ is the readout weight matrix and $\alpha \in (0, 1]$ is the leaking rate. ESNs allow us to avoid the aforementioned shortcomings of recurrent models by *fixing* input and recurrent transformation matrices $\mathbf{W}_{in}$ and $\hat{\mathbf{W}}$, and enforces stability of the dynamics by imposing the recurrent transformation to be asymptotically stable, i.e., $\rho(\hat{\mathbf{W}}) < 1$ [4]. As a result, the only set of free parameters is represented by readout weights $\mathbf{W}$.

One approach to learning the readout is to cast it to the problem of solving a linear system of the form $\mathbf{Y} = \mathbf{W}\mathbf{S}$, where $\mathbf{Y} \in \mathbb{R}^{N_y \times N}$ is the matrix containing the accumulated targets over N time steps, $\mathbf{S} \in \mathbb{R}^{N_x \times N}$ is the matrix composed by the reservoir states over time. This problem can be approached by minimization of the least squares, i.e. $\arg\min_{\mathbf{W}}(\mathbf{Y} - \mathbf{W}\mathbf{S})^2$, which can be easily solved via ridge regression, a closed-form solution formalized as $\mathbf{W} = \mathbf{Y}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \lambda\mathbf{I})^{-1}$, where $\lambda \in \mathbb{R}^+$ is the L2-regularization term and $\mathbf{I}$ is the identity matrix. Authors in [1] introduce an *exact* federated version of ridge regression for client-server topologies, namely `FedRR`. In this algorithm, each client $c \in \mathcal{C}$ computes the local matrices $\mathbf{A}_c = \mathbf{Y}_c\mathbf{S}_c^T$ and $\mathbf{B}_c = \mathbf{S}_c\mathbf{S}_c^T$. The server aggregates the matrices $\mathbf{A} = \sum_{c\in\mathcal{C}} \mathbf{A}_c$ and $\mathbf{B} = \left(\sum_{c\in\mathcal{C}} \mathbf{B}_c\right) + \lambda\mathbf{I}$ and computes $\mathbf{W} = \mathbf{A}\mathbf{B}^{-1}$.

Partial Federated Ridge Regression. We extend `FedRR` towards an approximate, communication-efficient solution, named Partial Federated Ridge Regression (`pFedRR`). Given an ESN with N_x recurrent units, each client c applies a policy to determine a set of indices $\mathcal{S}_c \subseteq \{0, 1, \dots, N_x - 1\}$ of the recurrent neurons whose information has to be forwarded to the server. Given the matrix $\mathbf{B}_c$ computed on the local data, the client sparsifies the matrix $\mathbf{B}_c$ as follows:

$$\tilde{\mathbf{B}}_c = \mathbf{B}_c \odot (\mathbf{M}_{\mathcal{S}} + I) \tag{2}$$

where $\mathbf{M}_{\mathcal{S}_c} \in \{0, 1\}^{N_x \times N_x}$ is a binary matrix such that $\mathbf{M}_{\mathcal{S}}[i, j] = 1$ if $i \neq j$, $i, j \in \mathcal{S}$, and I is the identity matrix. Then, the client sends the pair $(\mathbf{A}_c, \tilde{\mathbf{B}}_c)$

to the server, which aggregates the received matrices $\mathbf{A} = \sum_{c \in \mathcal{C}} \mathbf{A}_c$ and $\tilde{\mathbf{B}} = \sum_{c \in \mathcal{C}} \tilde{\mathbf{B}}_c$ and solves the approximate linear system as $\mathbf{W} = \mathbf{A}(\tilde{\mathbf{B}} + \lambda \mathbf{I})^{-1}$. Compressing only the matrix $\mathbf{B}_c$ is supported by the straightforward motivation that, from the communication perspective, the dimensionality of $\mathbf{B}_c$ is usually much larger than that of $\mathbf{A}_c$, since it is quadratic on the number of units N_x.

Importance-based Selection Policy. To comply with the requirement of `pFedRR`, we propose a selection policy where each client sends to the server the parameters which are most relevant with respect to their local distribution. For this purpose, we propose a method, namely Importance-based Partial Federated Ridge Regression (`I-pFedRR`) where the compression happens upon the selection of the most relevant neurons for the local task. To do so, we employ the Fisher Information Matrix (FIM), which estimates how much information a set of data provides about the unknown parameters of a statistical model. From now on, we omit the subscript denoting the k-th client for better clarity. Let's assume that our readout $f(\mathbf{W}) = \mathbf{W}\mathbf{S}_i + \epsilon$ is characterized by a measurement error ϵ sampled from a normal distribution. The log-likelihood of the data is given by the density:

$$\mathcal{L}(\mathbf{W}) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{-(\mathbf{Y}_i - \mathbf{W}\mathbf{S}_i)^2}{2\sigma^2}. \tag{3}$$

The FIM is defined as the expected value of the negative second derivative of the log-likelihood function, i.e.,

$$\mathcal{F}(\mathbf{W}) = -\mathbb{E}\left[\mathcal{H}(\mathbf{W})\right] = \mathbf{S}\mathbf{S}^T/\sigma^2. \tag{4}$$

For our purposes, the use of the FIM is to rank the neurons to determine which are most relevant for the local model. As a result, since the term σ is constant, we can neglect it from the formulation of the FIM and derive $\mathcal{F}(\mathbf{W}) = \mathbf{S}\mathbf{S}^T = \mathbf{B}$. From this formulation of the FIM, we define the *importance vector* for ranking the recurrent neurons as

$$\mathcal{I} = \sum_{j=0}^{N_x} \mathbf{B}_{ij}^2, \tag{5}$$

which is the sum over the columns of $\mathcal{F}$ squared. Then, we apply min-max rescaling to all the entries of $\mathcal{I}$ to have all the values in [0, 1]. Given this vector and a threshold hyperparameter $\tau \in (0, 1)$, we define the subset of selected neurons as $\mathcal{S} = \{i \,|\, 0 \leq i < N_x - 1; \mathcal{I}_i > (1 - \tau)\}$. This thresholding technique allows to select all the neurons whose local, relative importance is highest.

3 Experimental Assessment

The aim of our experiments is to assess the quality of the global readout when learned with the proposed technique. We assessed the effectiveness of `I-pFedRR` by comparing it to the baseline method `FedRR`, and `R-pFedRR` (i.e., `pFedRR` with random selection of the units for compression) on benchmarks with an increasing number of training clients, and increasing degree of compression. We conducted our experiments on two Human State Monitoring benchmarks: WESAD [5], a

dataset for stress and affect detection from wearable devices; HHAR [6], a dataset for activity recognition from inertial data. Both datasets lend themselves to the adaptation to a federated scenario thanks to the presence of IDs identifying both the user and the device. Being an activity recognition dataset with a low number of participants, HHAR represents a more challenging benchmark for the federated setting, since the heterogeneity of local client behaviors is not compensated by the presence of a sufficient number of participating clients.

Setup. The WESAD dataset was collected from 15 participants, each of which presents 8 synchronized time series of physiological data sampled at 700Hz by a chest-worn device. Each timestep in the sequence is equipped with a label denoting one of 4 cognitive states experienced during the collection. In the HHAR dataset, for each participant, we selected the sequence corresponding to the data from the LG Nexus4 smartphone, and downsampled it to 100Hz. Each timestep is associated with 6 features corresponding to intertial data, and with a label denoting one of the 6 activities performed by the participant. In both datasets, we chunked the sequence in sections of 700 and 200 timesteps for WESAD and HHAR respectively. Similarly to [2], we employed a client-wise training-validation-test split of the dataset, which is 9-3-3 and 5-2-2 for WESAD and HHAR respectively. We conducted our experiments by involving an incremental number of training clients, i.e., 25%, 50%, 75% and 100%.

Given a percentage of training clients, the experiments that we report were anticipated by a preliminary model selection to determine the best configuration of hyperparameters for an ESN with 1000 recurrent units. In particular, we evaluated configurations with $\rho \in [0.3, 0.99]$, input scaling in $[0.5, 1)$, leaking rate $\alpha \in [0.1, 0.9]$. In this phase, we selected the configuration with the highest validation accuracy, and fixed it for the subsequent experiments.

Then, for each percentage of training clients, we performed a comparison of the three methodologies with the following setup. First, we chose a value for τ for the current experiment. Then, the server instantiates a reservoir with the best configuration of the current percentage of training users and forwards it to the training clients. Each training client computes three pairs of matrices: one with `FedRR`; one with `I-pFedRR` with threshold τ; one with `R-pFedRR` with a chosen number of units equal to the one in `I-pFedRR`. Then, the server computes three readouts, one for each methodology. The resulting readouts are assessed on the validation clients. This experiment was conducted for all $\tau \in \{0.1, 0.2, \ldots, 0.9\}$, and repeated 5 times for each value of τ. The proposed setup ensures fairness across the methods, since all of them employ the same reservoir and the two selection policies in `pFedRR` select the same number of units for the compression.

Results. Table 1 reports a summary of the results in our experiments. For `FedRR`, we report the average test accuracy across all the experiments (since it is not subject to validation). For `I-pFedRR` and `R-pFedRR`, we report the average test accuracy and number of units chosen for compression with the τ whose validation accuracy across the 5 trials was highest. Figure 1 reports the behavior of `pFedRR` with both selection policies and increasing value of τ.

Users	WESAD				
	FedRR	I-pFedRR		R-pFedRR	
	Acc.	Acc.	% Units	Acc.	% Units
25%	72.41 ± 1.40	**80.71 ± 0.37**	**49.72 ± 1.71**	78.85 ± 1.04	91.28 ± 1.78
50%	73.56 ± 0.54	**77.74 ± 0.74**	**13.72 ± 1.01**	77.87 ± 1.05	90.86 ± 1.28
75%	75.50 ± 0.82	**79.77 ± 0.21**	**11.39 ± 1.14**	79.99 ± 0.43	89.71 ± 1.25
100%	76.92 ± 0.85	80.46 ± 0.32	1.22 ± 0.14	**80.55 ± 0.44**	**1.22 ± 0.14**

Users	HHAR				
	FedRR	I-pFedRR		R-pFedRR	
	Acc.	Acc.	% Units	Acc.	% Units
25%	72.95 ± 4.66	**68.73 ± 2.31**	**94.50 ± 0.81**	59.74 ± 1.69	94.50 ± 1.81
50%	76.31 ± 13.08	**57.93 ± 4.95**	**88.63 ± 1.32**	49.93 ± 3.21	88.63 ± 1.32
75%	84.96 ± 0.37	**69.83 ± 1.95**	**91.41 ± 1.06**	65.01 ± 1.56	91.41 ± 1.06
100%	83.53 ± 0.39	**74.39 ± 3.05**	**93.53 ± 1.76**	68.67 ± 3.21	93.53 ± 1.76

Table 1: Summary of the results. The best trade-off between test accuracy and % of units is reported in bold.

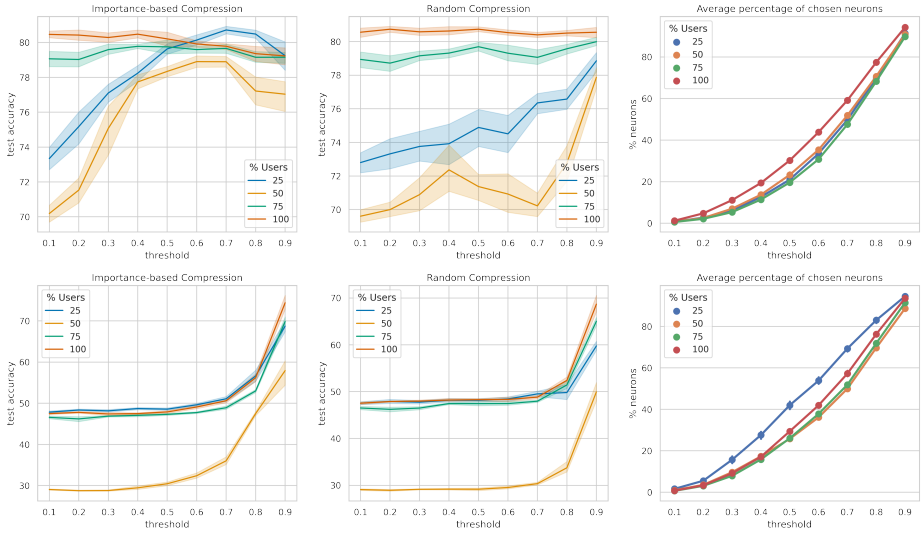

Fig. 1: Results for increasing value of τ on WESAD (above) and HHAR (below).

On WESAD, we can observe that both compression-based approaches outperform the baseline method `FedRR` with each percentage of training clients involved. This shows that in addition to the advantage of significantly reducing the communication cost, using a subset of the neurons for global readout learning also works as a regularization method that improves the generalization ability of the method. In addition, we can observe that with a lower number of training clients involved, the percentage of neurons chosen by `I-pFedRR` is significantly lower than `R-pFedRR`. Thus, in the presence of less information for readout learning, turning to a technique biased on relevant information is more performant. Finally, Figure 1 (above) shows that as the number of training clients increases, the importance-based method requires progressively fewer

units to achieve maximum accuracy on the test data. In particular, the cases with 25% and 50% training clients reach their peak accuracy with ∼50% and ∼13% of units, respectively.

On HHAR (see Table 1, and Figure 1 below), even if both compression-based methods achieve lower performance than the baseline, the importance-based method always yields the best trade-off between accuracy and communication efficiency. Finally, in Figure 1, the rightmost plots indicate that the growth of unit numbers is nearly linear for HHAR, but tends to be exponential for WESAD, suggesting that HHAR lacks a clear distinction between relevant and irrelevant information unlike WESAD.

4 Conclusions

We proposed a novel methodology for computing the Ridge Regression on Federated Echo State Networks with a communication-efficient approach. Each client selects a subset of the most relevant neurons and sparsifies the statistics to be forwarded to the server. We evaluated this methodology in comparison with the original method with full statistics, and one with random sparsification. The results show that using the importance policy in selecting the neurons allows for improving the generalization capabilities of the model while reducing the communication overhead.

In future work, we plan to develop a hybrid selection policy between importance and random to balance bias and variance in the information exchanged. In addition, this methodology can also be exploited in continual learning to encourage information sparsification in the global readout.

References

[1] Davide Bacciu et al. Federated reservoir computing neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2021.

[2] Valerio De Caro, Claudio Gallicchio, and Davide Bacciu. Federated adaptation of reservoirs via intrinsic plasticity. In *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, October 2022.

[3] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer science review*, 3(3):127–149, 2009.

[4] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

[5] Philip Schmidt et al. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[6] Allan Stisen et al. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140, 2015.