Subset Modelling: A Domain Partitioning Strategy for Data-efficient Machine-Learning

Vitor Ribeiro, Eduardo Pena, Raphael Saldanha, Reza Akbarinia, Patrick Valduriez, Falaah Arif, Julia Stoyanovich, Fabio Porto

Subset Modelling: A Domain Partitioning Strategy for Data-efficient Machine-Learning

Vitor Ribeiro[1], **Eduardo H. M. Pena**[4], **Raphael Saldanha**[3], **Reza Akbarinia**[3], **Patrick Valduriez**[3], **Falaah Arif Khan**[2], **Julia Stoyanovich**[2], **Fabio Porto**[1]

[1]National Laboratory for Scientific Computing (LNCC)

[2]Center for Data Science, New York University (NYU)

[3]Institut national de recherche en sciences et technologies du numérique (INRIA)

[4]Federal University of Technology (UTFPR)

victorr@posgrad.lncc.br, fporto@lncc.br, eduardopena@utfpr.edu.br

{raphael.de-freitas-saldanha, reza.akbarinia, patrick.valduriez}@inria.fr

{jds405, fa2161}@nyu.edu

Abstract. *The success of machine learning (ML) systems depends on data availability, volume, quality, and efficient computing resources. A challenge in this context is to reduce computational costs while maintaining adequate accuracy of the models. This paper presents a framework to address this challenge. The idea is to identify "subdomains" within the input space, train local models that produce better predictions for samples from that specific subdomain, instead of training a single global model on the full dataset. We experimentally evaluate our approach on two real-world datasets. Our results indicate that subset modelling (i) improves the predictive performance compared to a single global model and (ii) allows data-efficient training.*

1. Introduction

The remarkable success of machine learning systems (MLS) in various domains has been largely attributed to the availability of vast amounts of data and high-performance computing infrastructures [Zhang et al. 2022]. Such resources come at a significant cost, making them accessible mainly to larger organizations with substantial financial capabilities. A second challenge is that MLS do not perform uniformly well on all parts of the input space, despite showing good overall performance. This issue is pervasive across all MLS: disparities in performance across demographic subgroups have been the focus of fair machine learning [Chouldechova and Roth 2020].

This paper attempts to tackle both challenges simultaneously by considering "subdomains" within the input space. As a motivating example, consider the task of predicting the incidence of dengue in Brazil. In principle, we could build a single model that forecasts dengue prevalence in all Brazilian cities However, different cities experience variations in transmission patterns that may be hard to capture by a single model [Cabrera and et al 2022]. On the other hand, building a separate (local) model for each municipality is also challenging. First, there may be relatively few dengue cases in some areas, resulting in limited data for model training, and, thus, in models that will not generalize well when deployed. Second, training and maintaining separate models for each municipality demands significant data management effort and computing resources. The approach of this paper interpolates between these two extremes.

Summary of contributions. We propose a framework that accounts for shared characteristics and regional variations across different cities, notably at low training costs and good prediction capability. The idea is simple: first identify subsets within the dataset, and then train subset models. At inference time, assign the incoming sample to one of the subset, and use the corresponding model to make a prediction. It is important to mention that the clustering process may help us to achieve algorithmic fairness. We present our framework in Section 2, and show its effectiveness empirically in Section 3 using two tasks: dengue forecasting in Brazil and predicting unemployment in the US based on Census data.

2. Partition Modelling Framework

Preliminaries. Let a given dataset be denoted as $D(X, Y)$, where X represents the attribute set and Y is the target variable for the prediction task. We assume the existence of a function f such that $f(X) : X \rightarrow Y$. A learner in a machine learning context aims to construct a model $\hat{f}(X) : X \rightarrow \hat{Y}$ that can approximate f such that $f(X) \approx \hat{f}(X)$. The quality of the prediction is determined by an error metric μ such that $|f(x_i) - \hat{f}(x_i)| \leq \epsilon$, where ϵ is a real value computed according to μ. The empirical error of a model is computed over a set of samples in D according to the *Empirical Risk Minimization formulae* [Shalev-Shwartz and Ben-David 2014], whereas $E_s(\hat{f})$ represents an average measure over a set of observed input samples, meaning that the approximation $\hat{f}(X)$ can perform better for some examples than others.

$$E_s(\hat{f}) = \frac{|i \in [m] : \hat{f}(x_i) - f(x_i)|}{m}, [m] = 1, 2, ..., m, \forall x_1 \in X. \tag{1}$$

Our framework. We consider a dataset D whose domain can be one of two classes: time-series and independent samples. The first class D_r is interpreted as a set of spatially localized time-series (ts), ordered by time, and having some measurement values at each time instant. The second dataset D_c has a set of features X and a class label Y. The dataset D is composed of a set of samples $D = \{s_1, s_2, ..., s_n\}$. The problem we want to solve is to select partitions of D, $P_i = \{S_1, S_2, .., S_k\}$, with $S_j = \{s_{j1}, s_{j2}, .., s_{jn}\}$, $S_j \cap S_t = \emptyset$, if $j \neq t$, $D = \cup_{j=1..k} S_j, \forall P_i$, such that we train k machine learning models, using a given *learner* algorithm. Each machine-learning model m_j in M, $1 \leq j \leq k$, is built on the samples of the corresponding subset $S_j \in P_i$. We define P_i as a partition of dataset D. We want to find a partition P_i of the dataset D, such that when used for training produces models in M with higher accuracy than a global model trained on samples of the complete dataset D

$$P_i := argmin_{P_i \in P} \sum_{j=1..k} E_s((ts_l \in S_{ij}) : m_{ij}(ts_l)) \tag{2}$$

Solving equation 2 is not practical for large datasets due to the following combined reasons:(i) for D_r the number of windows computed on the pre-processing step for preparing the data for model training can considerably multiply the original dataset size, (ii) building and evaluating models for all subsets in all partitions (i.e. the power set of D_r) has complexity with lower bound exponential in the sizes of D_r.

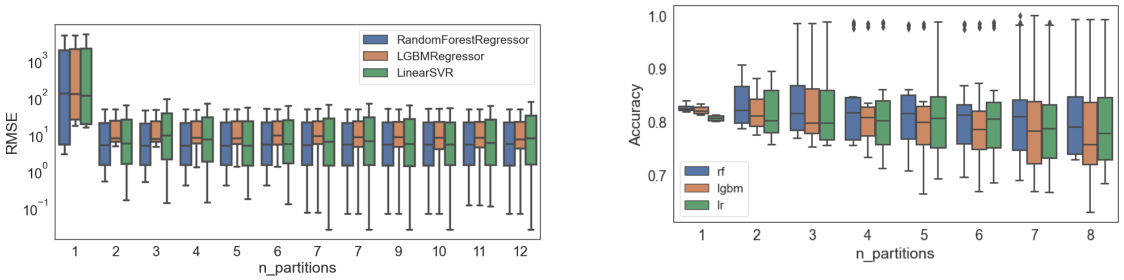
We consider that the partition of the dataset D we are looking for should strike a balance between similarity among the samples of a subset, and some variation that would enable the learner to generalize beyond observed trained data. In this paper, we argue that by adopting non-supervised clustering algorithms, such as *k-means* or *k-shape* the partition therein obtained offers the desired properties, approximating a solution for the Equation 2 And thus it becomes: $P_i = clustering\_algorithm(D, k) \;\; (3)$. Once the set of M models has been built, they are ready to be used for inference. The latter happens in the following way. Given an input sample s_x to which an inference is to be computed, we want to use a model $m_i \in M$ that is the best fit for s_x. The chosen model m_i is the one built on the samples of the partition S_i, whose representative is closest in a metric distance to s_x.

Related work. [Khan and Stoyanovich 2023] explore the existence of subdomains in the data through the lens of algorithmic fairness. Here, subdomains correspond to minority groups.The authors use this framing to motivate the use of group-specific models for improved performance on socially disadvantaged groups. In comparison, we compute subsets that offer no specific semantic guarantee; rather, they are identified based on data distributions similarity. The *Coreset* approach aims to select a subset of a dataset that produces models of comparable accuracy to a model built on the full dataset [Mirzasoleiman et al. 2020]. This work differs from ours because we aim at improving the model accuracy by training subset models on subsets of the original dataset.Our framework addresses the data subset selection problem. This problem is known to be NP-hard since both the training set and the model parameters work as optimization variables [Wei et al. 2015]. Still, our empirical results in Section 3 show that subset modeling works well in practice.

3. Experimental Evaluation

The *Dengue* dataset: The first task evaluates the subset modelling procedure on a time series regression task to forecast the number of weekly dengue virus (DENV) cases in each city in Brazil. DENV-suspected cases are tracked nationwide using a health information system. For this research, we computed the number of dengue cases each week by city from 2011 to 2020. We only considered confirmed cases. Further, due to the data sparsity, we only considered cities with more than 100,000 inhabitants. The dataset has 400 measurements of dengue cases in 312 cities in total.

The *ACSEmployment* dataset: The second task we consider is to predict a person's employment status, based on social and demographic information collected by the US Census. Folktables [Ding and et al 2021] is a benchmark dataset in fair machine learning, constructed from census data from 50 US states for the years 2014-2018. We report results on the ACSEmployment task: a binary classification task of predicting whether an individual is employed. We report our results on data from Georgia from 2018. The dataset has 16 features, including age, schooling, and disability status, and contains about 200k samples, which we sub-sample down to 10k samples for computational feasibility.



(a) RMSE on *Dengue*, lower is better.

(b) Accuracy on *ACSEmployment*, higher is better.

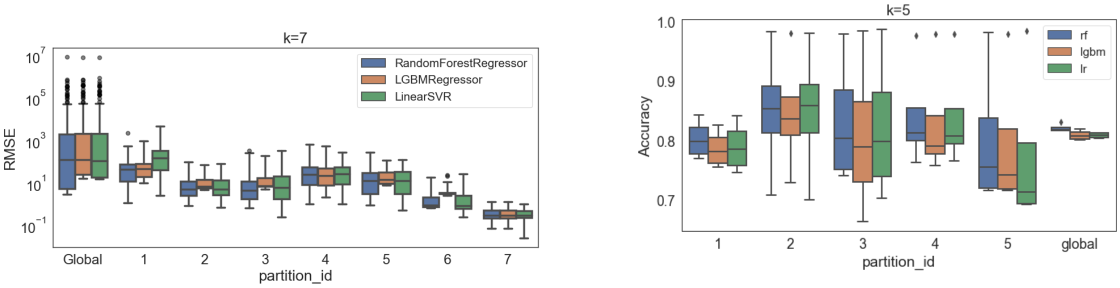
Figure 1. Performance of subset models for different learners; k is the number of partitions, $k = 1$ corresponds to the global model.

3.1. Experimental set-up

***Dengue*:** As a training feature, we considered the time series on the number of dengue cases. We computed a sliding window of length (w) and stride 1 on the complete training sequence, producing $n - (w + 1)$ subsequences as input for the training process. For each window of size w, the prediction infers the next measurement value. We first pre-processed the data to normalize it, and then split it into training and test, using the first (temporally) 320 measurements for training (80%) and testing on the remaining 80 (20%). We compared two training approaches: a single global model trained on all 312 municipalities vs. a subset-based approach. For the subset approach, we first partition the training dataset using *k-means* clustering for $k \in [4, 7]$, with dynamic time warping (DTW) as the distance metric. We then trained three learners: *RandomForestRegressor* (rf), *LGBMRegressor* (lgbm) and *LinearSVR* (lsvr), with hyperparameter tuning using Optuna. This procedure were repeated 100 times for each learner algorithm. During testing, for each municipality, we select for prediction the *subset model* trained on the subset containing the time-series for that municipality. We also use the same time-series as input for inference with the *global model*. Since the *Dengue* task is a time-series regression, we use the RMSE as the evaluation metric, and report results in Figures 1a and 2a.

***ACSEmployment*:** We used all 16 features as input for the binary classification task. We one-hot-encoded the categorical features, and performed standard scaling on the numerical ones. Next, we split the dataset into training and test sets (80:20). We ran k-means clustering using all the features, and partitioned the dataset into partitions based on cluster assignment. For each partition, we trained three learners: *Random Forest Classifier* (rf), *Gradient-boosting machine* (lgbm) and *Logistic Regression* (lr). We performed 3-fold cross validation using RandomizedSearchCV from scikit-learn to tune the hyperparameters of each subset model. We applied this methodology for $k \in [2, 8]$. To capture variance in the results, for each choice of k we repeated the experiment with 3 different random train-test splits. In each setting, we also trained a *global model* on the full dataset.

During testing, we first predict the cluster assignment for each test sample, and then use the corresponding *subset model* (trained on samples only from the same cluster) to do inference, making predictions for each point using the global model. We report accuracy as the evaluation metric for this binary classification task, in Figures 1b and 2b. To evaluate training efficiency, we report training dataset sizes for each subset model in Figure 3. For both experimental cases we decided to fix the learning algorithmns and their



(a) RMSE for *Dengue*, $k = 7$, lower is better.

(b) Accuracy for *ACSEmployment*, $k = 5$, higher is better.

Figure 2. Performance of subset models vs. the global model, for a fixed k.

hyper-parameters, trying to establish a simpler correlation between cluster composition and prediction quality.

3.2. Results

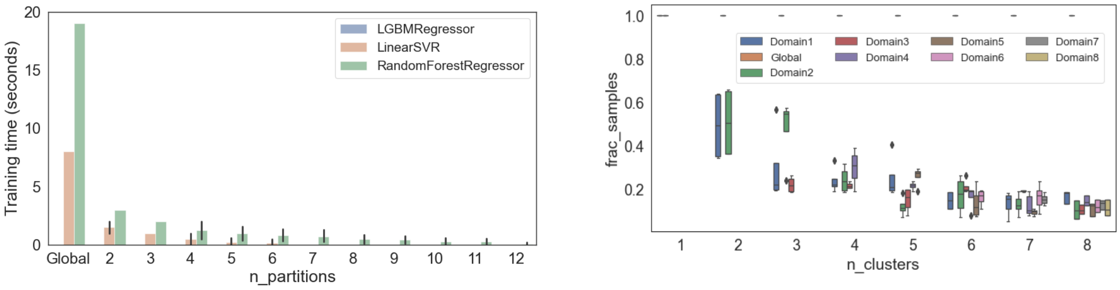
Figure 1 provides evidence that partitioning the data to guide the construction of subset models can improve (lower) RMSE. For any choice of $k > 2$, we see an improvement in performance on *Dengue*, in terms of lower root mean square error (RMSE). For *ACSEmployment*, we see some improvement over the global model, for some values of k (e.g., $k = 2$), in terms of higher accuracy. For both tasks, as the number of partitions becomes larger, the variance of the accuracy metric (shown by the width of the box and whiskers plot) increases. This is expected, since the number of training samples per cluster will decrease, leading to higher variance. This results may indicate the need for considering some constraint regarding the classification task.

In Figure 2, we present performance for a fixed k, with $k = 7$ for *Dengue* and $k = 5$ for *ACSEmployment*. We see that all of subset models outperform the global model on *Dengue* (lower RMSE). On *ACSEmployment*, some subset models (such as the *rf* and *lr* on partition 2) outperform the global models, indicating improved performance on some subset of the input space. Figure 3 ilustrates the training time and dataset size for different subset models, as a function of k, to evaluate the efficiency gains from this methodology. Figure 3a,Presentst the elapsed time as the maximum time to train a subset model for *Dengue*. Note that we only considered time dedicated with training time, disregarding clustering time.

Observe that, for all values of k, training time decreases as the number of partitions increases, reinforcing the possibility that *subset models* can expedite the training process. Further, note that considering each partition as a disjoint dataset brings the training problem a natural parallelism. Figure 3b reports the fraction of samples assigned to each cluster for *ACSEmployment*. As expected, larger number of partitions decreases the fraction of samples per subset model. This causes a larger variance in performance of subset models for $k > 7$ on both tasks, indicating the existence of a trade-off between performance and efficiency.

4. Conclusion

The machine learning literature provides several pieces of evidence indicating that data selection can have an impact on training time while maintaining prediction quality. Typ-



(a) Training time on *Dengue*.

(b) Fraction of samples per subset for *ACSEmployment*.

Figure 3. Computational efficiency, for different k (number of clusters/domains), measured by training time for *Dengue* (a) and dataset fraction for *ACSEmployment* (b).

ically, this techniques support the construction of a single model. In this paper, we propose a novel training approach that extends the data selection problem, providing support for constructing multiple models. Our experimental findings suggest that a subset approach can improve predictive performance, as well as training efficiency, bounded by an accuracy-efficiency trade-off.

References

Cabrera, M. and et al (2022). Dengue prediction in latin america using machine learning and the one health perspective: A literature review. *Tropical Medicine and Infectious Disease*, 7(10):322.

Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89.

Ding, F. and et al (2021). Retiring adult: New datasets for fair machine learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *NeurIPS 2021, December 6-14, 2021, virtual*, pages 6478–6490.

Khan, F. A. and Stoyanovich, J. (2023). The unbearable weight of massive privilege: Revisiting bias-variance trade-offs in the context of fair prediction. *arXiv preprint arXiv:2302.08704*.

Mirzasoleiman, B., Bilmes, J. A., and Leskovec, J. (2020). Coresets for data-efficient training of machine learning models. In *ICML 2020, 13-18 July 2020*, volume 119, pages 6950–6960. PMLR.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Wei, K., Iyer, R., and Bilmes, J. (2015). Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1954–1963. JMLR.org.

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., and Perrault, R. (2022). The ai index 2022 annual report.