

Symbolic Regression for Medical Scoring Systems: a Bayesian and Multi-Objective Approach

Mattia Billa^{1,*}

¹*Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Univ. Modena e Reggio Emilia, Modena, Italy*

Abstract

Scoring systems play an important role in high-stakes domains, such as medicine, by quantifying complex phenomena through the combination of data features, thereby assisting decision-making processes and clinical research. Traditional methods often rely on linear models, which may struggle to capture the complexity inherent in data. Recently, Symbolic Regression has emerged as a promising alternative, offering the ability to construct nonlinear models that are both interpretable and accurate. However, this approach faces some limitations, including a lack of uncertainty awareness and difficulties in adapting to non-IID scenarios such as those found in Federated and Continual Learning settings.

We propose a novel data-driven approach that integrates Symbolic Regression with Bayesian Inference and Multi-Objective Optimization. By combining these methodologies, our approach aims to address both uncertainty quantification and adaptability in Continual and Federated Learning scenarios. Initial experiments on clinical data have shown promising results, highlighting the potential of the proposed framework for improving the reliability and applicability of scoring systems in medical contexts.

Keywords

Scoring Systems, Symbolic Regression, Federated Learning, Continual Learning, Bayesian Inference, Multi-Objective Optimization

1. Introduction

The integration of Artificial Intelligence (AI) and medicine has given rise to an innovative approach known as P4 medicine. P4 medicine combines predictive, preventive, personalized, and participatory healthcare, presenting a paradigm shift in healthcare delivery [1]. Data from Electronic Health Records (EHRs), Patient Reported Outcomes (PROs), and wearable devices, combined with statistical and machine learning methods, have significant potential in clinical research, decision support, and knowledge discovery.


However, data-driven healthcare applications present unique challenges due to the constantly evolving and often poorly controlled environment in which they are developed. Clinical data, for example, is highly sensitive, expensive to acquire and curate, and subject to complex governance policies. Moreover, medical data is often distributed across multiple healthcare institutions and facilities. In specific scenarios, combining knowledge from multiple institutions becomes necessary to improve model performance, overcome data scarcity, or validate results. Nevertheless, this should be done while safeguarding patient privacy and data security. A possible solution to address these challenges is Federated Learning [2]. This approach allows

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

✉ mattia.billa@unimore.it (M. Billa)

ORCID [0009-0005-1979-8918](https://orcid.org/0009-0005-1979-8918) (M. Billa)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

models to be trained across decentralized devices or servers without exchanging raw data, thereby preserving the confidentiality of sensitive information.

Overcoming these obstacles, while maintaining model interpretability, is crucial to ensure the successful integration of AI in healthcare and to maximize its potential for improving patient care and outcomes [3]. Recent machine learning methods for interpretable scoring systems mostly employ linear classification models [4], hence assuming a fixed index form. An alternative approach is Symbolic Regression [5], a technique aiming to discover mathematical expressions approximating a dataset without relying on predefined functional forms.

The primary objective of this contribution is to address both result interpretability and data distribution issues for the development of nonlinear scoring systems. We will accomplish this by combining Symbolic Regression, Multi-Objective Optimization [6], and parametric Bayesian Inference. Indeed, since the current approach to Symbolic Regression doesn't take into account the epistemic and aleatoric uncertainty, we propose to integrate Bayesian Inference [7] into our framework. We also plan to exploit the uncertainty quantification, together with Multi-Objective optimization, to address Continual and Federated Learning in non-IID scenarios [8].

The following sections are structured as follows: Section 2 provides an overview of related and background material; Section 3 describes the proposed approach; Section 4 presents some preliminary results; Lastly, Section 5 concludes with final remarks.

2. Background and related work

2.1. Scoring systems

Scoring systems are mathematical equations that combine elementary indicators to describe complex phenomena with a single value, providing decision-support tools. Examples in clinical settings include BMI and Charlson score [9]. Traditionally, domain experts have developed these scores using trial and error methods. However, current efforts concentrate on data-driven approaches, emphasizing the importance of interpretability in the generated models.

Symbolic Regression is a potential solution to this end, consisting of finding a mathematical expression that best fits a given dataset without assuming a specific form beforehand [5]. This problem is usually solved using Genetic Programming (GP) [10], i.e. an evolutionary approach that encodes mathematical formulas as unary/binary trees. The main goal of GP is to select, simulating a natural selection process, the model that optimizes a particular loss function over a dataset. In the context of scoring system development, where data can be small and unbalanced, we are also interested in data sample stratification and balancing, aside from accuracy. Therefore, previous work was focused on tackling these issues as a Multi-Objective Optimization problem (MOO) [11, 6], using the NSGA-II evolutionary algorithm [12].

2.2. Learning with non-IID data

Federated Learning Growing concerns about data privacy and the need to combine knowledge from different facilities have led to the introduction of Federated Learning (FL). The goal of FL is to train a joint model in a decentralized way using data distributed across multiple devices [13]. FedAvg [14] was the first FL approach able to achieve good performance on distributed

datasets assuming IID data. However, in real-world scenarios, data is usually heterogeneous, with different statistical distributions from each device. Therefore, subsequent research has investigated the convergence on non-IID data [15] and has proposed a regularized approach for heterogeneous networks [16].

Other approaches, such as client or group personalization, have been introduced. These approaches allow individual devices, or similar groups of devices, to acquire a personalized model. Additionally, techniques like Domain Transformation and Domain Adaptation aim to either transform the perceived input space into a common input space or measure dissimilarities between datasets, allowing for the adjustment of the training model accordingly [8].

Recently, even Bayesian Learning has been considered in FL settings [17], taking advantage of its uncertainty quantification and performance on limited and heterogeneous data. For example, pFedBayes [18] is a personalization approach that uses the global distribution as a prior distribution and tries to minimize the KL divergence. In [19], online Laplace Approximation is used to approximate the local and global posterior, reducing the aggregation error. Instead, FedBE [20] is based on Bayesian model Ensemble to perform the aggregation step, achieving good performance on non-IID data.

Continual Learning In scenarios characterized by dynamic data collection and ongoing updates to datasets, data distribution may change over time, a phenomenon referred to as *concept drift*. This kind of time-wise heterogeneity is tackled by the Continual Learning (CL) paradigm, whose main challenge is the so-called *catastrophic forgetting*, i.e. the forgetting of previously learned concepts.

We can identify two stages of learning under concept drift, the first one is the Drift Detection. The goal is to understand whether a drift has occurred [8], and this can be done with Data Distribution-based methods, which use the statistical properties of data distributions, or Error Rate-based, which are based on the accuracy, or the uncertainty, of the model through time. The other stage is the Drift Adaptation, preventing the model from decreasing accuracy on new data, while not forgetting the previous data. Two main strategies are used for Drift Adaptation: memory-based methods [21], also known as rehearsal, and regularization methods [22].

Other approaches to CL rely on Bayesian Learning. For example, [23] employs online variational inference, using the previous posterior distribution as the new prior and multiplying it with the likelihood of the new data. Similarly, [24] uses the uncertainty of the parameters, obtained through Bayesian Inference, to change their learning rates.

Only a few works considered both Federated and Continual learning settings, such as [25], which performs both Drift Detection and Adaptation using the uncertainty related to the classifier on a sliding window and storing samples in a long-term memory.

While most of these works rely on neural networks, little research has yet been done in the Symbolic Regression field in the context of Continual and Federated learning. For instance, [26] proposes a federated Genetic Programming framework based on the aggregation of the local fitness (the loss of the model), achieving better generalization performance compared to models trained only on local datasets. However, this approach doesn't take into account the uncertainty related to the model, the relative importance that each dataset can have, and the possibility of updating the model after receiving more data.

3. A Bayesian and Multi-Objective approach to Symbolic Regression in non-IID scenarios

In Section 1, some of the limitations of previous work concerning Symbolic Regression have been introduced. The current multi-objective framework (MOSR) for scoring system development [6] doesn't take into account the uncertainty of the models, it lacks a formal mechanism for incorporating prior knowledge or updating the current one. To tackle these issues, we introduce an extension of this framework based on Bayesian Inference.

The key idea behind the Bayesian extension of MOSR is to replace the numerical constants inside each model with random variables whose initial distribution will encode prior knowledge, turning each model, estimated through maximum likelihood estimation (MLE), into a Bayesian model. The inference process, based on a Markov Chain Monte Carlo (MCMC) algorithm [7], should then return a posterior distribution of the model's parameters that also incorporates the uncertainty of the model itself. For the sake of simplicity, let's direct our attention towards a single model from this point onward. The standard assumption behind SR is that models are normally distributed around their expected values, given by a nonlinear expression $f(x, \theta)$ involving parameters θ , and with a standard deviation of σ .

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \text{ with } \mu_i = f(X_i, \theta)$$

As new data is acquired, we can update the prior distribution of θ by simply using its previous approximated posterior distribution, iteratively. According to the Bayes rule, a batch update of the posterior is equivalent to a sequential update of the posterior:

$$p(\theta|\mathcal{D}_{1:n}) = \frac{p(\theta)p(\mathcal{D}_{1:n-1}|\theta) p(\mathcal{D}_n|\theta)}{p(\mathcal{D}_{1:n-1}) p(\mathcal{D}_n)}$$

where \mathcal{D}_t is the subset of data at time t , $p(\theta)$ is the prior of θ , $p(\mathcal{D}|\theta)$ is the likelihood, and $p(\theta|\mathcal{D})$ is the posterior. The results of this approach are shown in the next section.

The proposed framework also aims to deal with Federated Learning scenarios by using Multi-Objective Optimization (MOO). The goal of MOO is to optimize problems with more objective functions that may conflict with each other. In this case, each objective function represents the loss function, such as the BIC criterion, of the considered model on a local dataset. Therefore, the current MOSR framework is naturally extended using the evolutionary algorithm to optimize the models on more datasets rather than different objective functions.

In MOO, a recurring issue is the presence of non-comparable objective functions. This means that the objectives cannot be directly compared in a meaningful way. In realistic Federated Learning settings, datasets from each client can differ in size, changing the magnitude of the corresponding loss functions. To make the posterior distributions comparable, we plan to use a fractional likelihood, introducing a *temperature* parameter.

So, at each generation, the server generates a population of models, which are then sent to the clients. Each client computes the loss of each model on its local dataset and then sends the evaluation back to the server. The server uses the received evaluations to select and evolve the population of models, optimizing the loss functions in a MOO manner. At the end of the process, the parameters of the models are inferred using the sequential procedure described for

Table 1
Results on the Body Fat dataset.

| Sampler | # samples | W_50 | W_95 | ESS bulk | ESS tail | R hat | Time (s) |
|---------|-----------|------|------|----------|----------|-------|----------|
| NUTS | 2000 | 39.1 | 95.6 | 2038.6 | 1722 | 1.00 | 89 |
| | 1000 | 36.9 | 95.6 | 1099.6 | 802.2 | 1.00 | 59 |
| HMC | 2000 | 41.3 | 95.6 | 1820.8 | 2002.2 | 1.00 | 39 |
| | 1000 | 36.9 | 95.6 | 889.6 | 1067.8 | 1.00 | 28 |
| MH | 10000 | 39.1 | 95.6 | 416.0 | 465.4 | 1.012 | 57 |
| | 1000 | 36.9 | 95.6 | 42.0 | 63.4 | 1.084 | 10 |

CL, which is equivalent to estimating parameters in a centralized scenario. Importantly, no raw data is transmitted over the network, maintaining privacy and security.

4. Preliminary results

The Bayesian extension for MOSR has been tested on both real-world and synthetic datasets. The real-world data was sourced from both a publicly available repository and the private Electronic Health Record (EHR) system. To be concise while maintaining generality, we present the results just on the Body Fat Percentage dataset¹, which contains body fat percentage estimates based on Siri’s formula, along with 14 anthropometric measurements of 252 men. Underwater weighing and body fat percentage were removed from the dataset because the former is not easy to obtain, and we want to compare the latter with the solutions coming from the MOSR.

The results are reported in Table 1, comparing different MCMC sampling algorithms (NUTS, Hamiltonian Monte Carlo, and Metropolis-Hastings) and sampling sizes. To test the uncertainty quantification of the framework, we introduce two metrics: *within 50* (W_50) and *within 95* (W_95). The within 50/95 metric quantifies the proportion of observed values that fall within their 50/95% posterior prediction interval, respectively. We have also included some converge metrics, such as the Effective Sample Size (ESS) and the \hat{R} . The results suggest that the framework can capture the uncertainty within the 95% prediction intervals but tends to be overconfident within the 50% prediction intervals. Concerning the convergence, NUTS and HMC sampling outperformed MH, even with fewer samples and comparable training times.

To assess the framework in a continual learning scenario, we split the Body Fat dataset into two distinct populations. The model is first trained on the first population, using the NUTS algorithm with two chains of 1000 samples. Then the posterior is sequentially updated using just the other population data. The results are shown in Table 2.

Not only did the predictive performance of the model not decrease, but it appears that the integration of the second population also improved the regression performance and posterior predictive intervals, especially for the 50% interval. So, this extension seems not to suffer from catastrophic forgetting.

¹<https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset>

Table 2

Results on the Body Fat dataset, integrating the second population.

| Population | W_50 | W_95 | ESS bulk | ESS tail | R hat | Time (s) |
|------------|------|------|----------|----------|-------|----------|
| 1 | 39.1 | 93.4 | 1099.6 | 802.2 | 1.00 | 110 |
| 1 and 2 | 43.4 | 1.00 | 822.4 | 977.8 | 1.00 | 190 |

5. Discussion and conclusion

In this contribution, we have addressed the challenge of building interpretable scoring systems in real-world scenarios characterized by distributed and evolving data. Some selected works about Federated and Continual Learning have been reported, with a focus on Bayesian approaches, which exhibit relevant properties related to uncertainty quantification.

Our proposed approach aims to integrate Symbolic Regression with parametric Bayesian Inference and MOO, ensuring both interpretability and data privacy. Initial testing on clinical data in a Continual Learning setup has shown promising results, demonstrating the potentiality of our framework in this kind of setting. In addition, a Federated Learning strategy that makes use of both Bayesian Learning and MOO has been introduced.

In future work, we plan to investigate the performance of this framework under both Continual and Federated Learning settings, simulating different degrees of heterogeneity of data between clients. Furthermore, scalability concerns will be addressed to ensure the efficacy of our approach as the number of clients and the amount of data increase.

References

- [1] M. Flores, G. Glusman, K. Brogaard, N. D. Price, L. Hood, P4 medicine: how systems medicine will transform the healthcare sector and society, *Personalized Medicine* 10 (2013) 565–576. doi:10.2217/pme.13.57.
- [2] J. Konečný, H. B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence, 2016. doi:10.48550/ARXIV.1610.02527.
- [3] F. Mandreoli, D. Ferrari, V. Guidetti, F. Motta, P. Missier, Real-world data mining meets clinical practice: Research challenges and perspective, *Frontiers in Big Data* 5 (2022). doi:10.3389/fdata.2022.1021621.
- [4] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, *Machine Learning* 102 (2016) 349–391.
- [5] W. La Cava, P. Orzechowski, B. Burlacu, F. O. de França, M. Virgolin, Y. Jin, M. Kommenda, J. H. Moore, Contemporary symbolic regression methods and their relative performance, 2021. doi:10.48550/ARXIV.2107.14351.
- [6] D. Ferrari, V. Guidetti, F. Mandreoli, Multi-objective symbolic regression for data-driven scoring system management, in: *2022 IEEE International Conference on Data Mining (ICDM)*, 2022, pp. 945–950. doi:10.1109/ICDM54844.2022.00112.
- [7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian data analysis*, CRC press, 2013.

- [8] M. F. Criado, F. E. Casado, R. Iglesias, C. V. Regueiro, S. Barro, Non-iid data and continual learning processes in federated learning: A long road ahead, *Information Fusion* 88 (2022).
- [9] M. E. Charlson, P. Pompei, K. L. Ales, C. MacKenzie, A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation, *Journal of Chronic Diseases* 40 (1987) 373–383. doi:10.1016/0021-9681(87)90171-8.
- [10] J. R. Koza, Genetic programming as a means for programming computers by natural selection, *Statistics and computing* 4 (1994) 87–112.
- [11] J. Kubalík, E. Derner, R. Babuška, Symbolic regression driven by training data and prior knowledge, in: *Proc. of the 2020 Genetic and Evolutionary Computation Conference*, 2020.
- [12] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, *IEEE transactions on evolutionary computation* 6 (2002) 182–197.
- [13] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (2019) 1–19.
- [14] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.
- [15] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of fedavg on non-iid data, *arXiv preprint arXiv:1907.02189* (2019).
- [16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, *Proceedings of Machine learning and systems* 2 (2020) 429–450.
- [17] L. Cao, H. Chen, X. Fan, J. Gama, Y.-S. Ong, V. Kumar, Bayesian federated learning: A survey, *arXiv preprint arXiv:2304.13267* (2023).
- [18] X. Zhang, Y. Li, W. Li, K. Guo, Y. Shao, Personalized federated learning via variational bayesian inference, in: *International Conference on Machine Learning*, 2022.
- [19] L. Liu, X. Jiang, F. Zheng, H. Chen, G.-J. Qi, H. Huang, L. Shao, A bayesian federated learning framework with online laplace approximation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2024) 1–16. doi:10.1109/TPAMI.2023.3322743.
- [20] H.-Y. Chen, W.-L. Chao, Fedbe: Making bayesian model ensemble applicable to federated learning, *arXiv preprint arXiv:2009.01974* (2020).
- [21] A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal, *Connection Science* 7 (1995) 123–146.
- [22] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, *Proceedings of the national academy of sciences* 114 (2017) 3521–3526.
- [23] C. V. Nguyen, Y. Li, T. D. Bui, R. E. Turner, Variational continual learning, *arXiv preprint arXiv:1710.10628* (2017).
- [24] S. Ebrahimi, M. Elhoseiny, T. Darrell, M. Rohrbach, Uncertainty-guided continual learning with bayesian neural networks, *arXiv preprint arXiv:1906.02425* (2019).
- [25] F. E. Casado, D. Lema, M. F. Criado, R. Iglesias, C. V. Regueiro, S. Barro, Concept drift detection and adaptation for federated and continual learning, *Multimedia Tools and Applications* (2022) 1–23.
- [26] J. Dong, J. Zhong, W.-N. Chen, J. Zhang, An efficient federated genetic programming framework for symbolic regression, *IEEE Transactions on Emerging Topics in Computational Intelligence* (2022).