Expert Selection in Distributed Gaussian Processes: A Multi-label Classification Approach

Hamed Jalali
Department of Computer Science
University of Tübingen
72076 Tübingen, Germany
hamed.jalali@wsii.uni-tuebingen.de

Gjergji Kasneci
Department of Computer Science
University of Tübingen
72076 Tübingen, Germany
gjergji.kasneci@uni-tuebingen.de

Abstract

By distributing the training process, local approximation reduces the cost of the standard Gaussian Process. An ensemble technique combines local predictions from Gaussian experts trained on different partitions of the data by assuming a perfect diversity of local predictors. Although it keeps the aggregation tractable, this assumption is often violated in practice. Taking dependencies between experts enables ensemble methods to provide consistent results. However, they have a high computational cost, which is cubic in the number of experts involved. By implementing an expert selection strategy, the final aggregation step uses fewer experts and is more efficient. Indeed, a static selection approach that assigns a fixed set of experts to each new data point cannot encode the specific properties of each unique data point. This paper proposes a flexible expert selection approach based on the characteristics of entry data points. To this end, we investigate the selection task as a multi-label classification problem where the experts define labels, and each entry point is assigned to some experts. The proposed solution's prediction quality, efficiency, and asymptotic properties are discussed in detail. We demonstrate the efficacy of our method through extensive numerical experiments using synthetic and real-world data sets.

1 Introduction and Related Works

Gaussian processes (GPs) [1] are interpretable and powerful Bayesian non-parametric methods for non-linear regression. By applying Bayes' theorem for inference, the posterior predictive distribution of a GP is the best linear unbiased estimator (BLUE) under the assumed model and provides proper quantification of the prediction error uncertainty. GPs do not need restrictive assumptions of the model and can estimate complex linear and non-linear structures. While GPs are extensively used in practical cases [2, 3, 4, 5, 6, 7], their cubic training and quadratic prediction costs [1] limit their application to big data use cases [8].

To reduce computational expense, the prominent distributed Gaussian processes (DGPs) divide the training data into numerous subsets. In this scenario, the local inference is conducted for each partition separately, and then the local estimations are combined through ensemble learning [9, 10, 11]. A local GP with expertise in a particular section is called an expert. Experts share the same hyper-parameters, thus accounting for implicit regularisation and encountering overfitting [8, 12]. In a DGP, the *conditional independence* (CI) assumption between partitions (i.e., between experts given the target) allows for factorizing the global posterior distribution as a product of local distributions. While this assumption reduces the computational cost, it results in inconsistencies and suboptimal

[1] I.e., in the size of the training set.

Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems (NeurIPS 2022).

solutions [13]. Relaxing the independence assumption raises the aggregation's theoretical properties. If the experts' predictions are assumed to be random variables, their relative correlations define dependencies between experts. The aggregated posterior distribution, in this case, provides high-quality forecasts and is capable of returning consistent results [14]. However, solutions that deal with the consistency problem suffer from extra computational costs induced by the need to find the inverse of the covariance matrix between experts for each test point [12].

To boost the efficiency of dependency-based aggregation, authors in [15, 16] discuss complexity reduction as an expert selection scenario that excludes a subset of original experts and considers only the valuable experts in the aggregation. For this purpose, they use the Gaussian graphical model (GGM), where the nodes are original experts, and the edges are their interactions. The nodes with fewer interactions are defined as unimportant experts and excluded from the model. This approach can lower the complexity and provide a good approximation for the original estimator. However, it is not flexible concerning new entries, and the selected experts are fixed for all test points.

The critical contribution of our work lies in selecting a dynamic and entry-dependent subset of local experts for each new data point using multi-label classification (MLC). We propose a flexible mechanism for each new observation that designates related experts to provide local predictions. Multi-label classification [17] is a generalization of multi-class classification, where multiple labels may be assigned to each instance. To transform the distributed learning case into a multi-label classification, the indices of the partitions/experts are the labels/classes. The task is to assign some experts to a new data point. Multi-class classification problem that allocates only one expert per test point produces discontinuous separation boundaries between sub-regions and therefore is not a proper solution for quantifying uncertainties [18, 12].

In this work, we adopt two MLC algorithms to assign experts to data points: k-nearest neighbors (KNN) and deep neural networks (DNN). For the first one, we use the partition's centroid as a substitute for the corresponding local expert. By estimating the distance between a new entry point and the centroids, we can find its K nearest neighboring experts. For the second approach, we train a neural network with a soft-max output layer and log-loss (i.e., cross-entropy loss) using the train points and their related partition indices that shows the partition they belong to. After training the DNN, we send a new test point through the network, and the experts with higher probability are assigned to this test point. Relative to consistent aggregation methods that use dependency information, our approach keeps all asymptotic properties of the original baseline and substantially provides competitive prediction performance while leading to better computational costs than other SOTA approaches, which use the dependency assumption.

The structure of the paper is as follows. Section 2 introduces the problem formulation and related work. In Section 3, the proposed model and the inference process are presented. Section 4 shows the experimental results, and we conclude in Section 5.

2 Problem Set-up

Background: We consider the non-linear regression problem $y = f(x) + \epsilon$, where $x \in R^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and the Gaussian likelihood is $p(y|f) = \mathcal{N}(f, \sigma^2 I)$. The objective is to learn the latent function f from a training set $\mathcal{D} = \{X, y\}$ of size n. The GP then describes a prior distribution over the latent functions as $f \sim GP\,(0, k(x, x'))$, where $k(x, x')$ is the covariate function (kernel) with hyperparameters ψ, and $x, x' \in X$. The training step of GPs scales as $\mathcal{O}(n^3)$ because it is affected by the inversion and determinant of the $n \times n$ matrix $\mathcal{C}$. It imposes limitations on the scalability of GPs for large data sets. In distributed GP (DGP), one divides the full training data set into M partitions (called experts) $\mathcal{D}^{'} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$ and trains standard GPs on these partitions. All GP experts are trained jointly and share a single set of hyper-parameters $\theta = \{\sigma^2, \psi\}$. For a test set X^* of size n_t, the predictive distribution of the i-th expert $\mathcal{M}_i$ is $p_i(y^*|\mathcal{D}_i, X^*) \sim \mathcal{N}(\mu_i^*, \Sigma_i^*)$.

Aggregation: Aggregating the experts in a DGP is typically based on the assumption that they are conditionally independent (CI). Therefore, for a test input x^*, the posterior distribution of DGP is given as:

$$p(y^*|\mathcal{D}, x^*) \propto \prod_{i=1}^{M} p_i^{\beta_i}(y^*|\mathcal{D}_i, x^*), \tag{1}$$

where the weights $\beta = \{\beta_1, \ldots, \beta_M\}$ describe the importance and influence of the experts. The most popular aggregation methods are the product of experts [19, 20] and Bayesian committee machine [21, 8, 22]. However, their predictions are not accurate enough, and CI-based aggregation generally returns a sub-optimal solution [23].

To consider dependencies between experts in the aggregation step, assume that the means of local predictive distributions μ_i^* are random variables and $\mu^*(x^*) = [\mu_1^*(x^*), \ldots, \mu_M^*(x^*)]^T$ be an $M \times 1$ vector that contains the centered predictions of M experts for a given test point $x^* \in X^*$. This allows us to consider correlations between the experts' predictions and latent variable y^*, $k_A(x^*) = Cov\left(\mu^*(x^*), y^*(x^*)\right)$, and also leverage internal correlations between experts, $K_A(x^*) = Cov\left(\mu^*(x^*), \mu^*(x^*)\right)$. Thus, for the test point, x^* and sub-model predictions $\mu^*(x^*)$, the aggregated predictor is defined as

$$y_A^*(x^*) = k_A(x^*)^T K_A(x^*)^{-1} \mu^*(x^*). \tag{2}$$

In [14], the authors showed that this linear estimator is the *best linear unbiased predictor* (BLUP). This method is known as the nested pointwise aggregation of experts (NPAE) and provides high-quality predictions. However, it needs the inverse of the $M \times M$ auto-covariance matrix $K_A(x^*)$, and its time complexity is cubic in the number of experts, i.e., $\mathcal{O}(n_t M^3)$, and it is not an efficient solution for large real-world data sets.

3 Expert Selection in Local Approximation GPs

Despite the high aggregation accuracy in 2, its computational cost is a challenge for using the method on large data sets. Expert selections can improve the performance of dependency-based aggregation in two ways. First, unrelated experts for a given data point can be excluded, and only informative partitions can be considered to make the prediction. Second, mitigating the number of experts reduces the prediction cost and enables the ensemble to be used in large data sets.

3.1 Expert Selection Using Graphical Models

Gaussian graphical model (GGM) is the continuous form of pairwise Markov random fields. It assumes the nodes of an undirected graph are random variables, and the joint distribution of the random variables is multivariate Gaussian distribution with zero mean and precision matrix Ω, $\mathcal{N}(0, \Omega^{-1})$. The elements of the precision matrix are the unknown parameters and show interactions between experts (edges in the graph). In the DGP framework, the experts' predictions μ^* are the nodes in a GGM. Authors in [15] proposed an expert selection method that uses only the highly connected nodes in the graph as the most critical experts in final aggregation defined in Equation 2. Although this method can mitigate the computational cost, it selects a fixed and static set of experts for all test points and, therefore, can not capture the specific properties of the new data points.

3.2 Multi-label Classification for Flexible Expert Selection

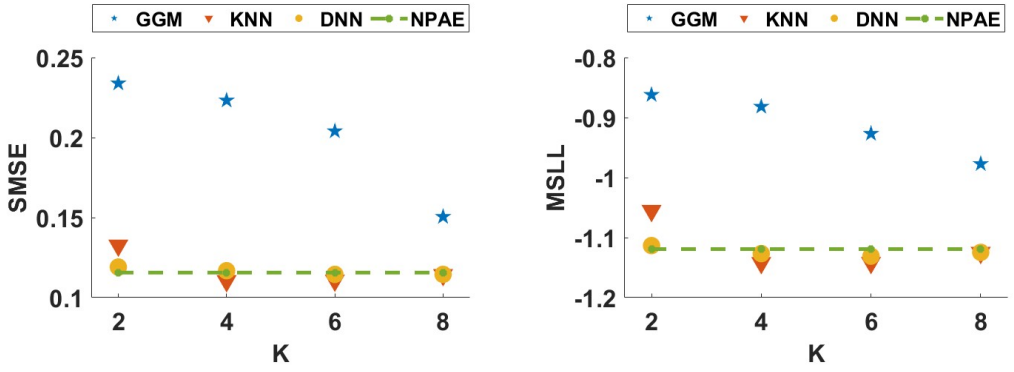
Assigning experts to new entry points in a distributed learning model can be a classification problem. Let's assume that each expert is a class of estimators. The selection problem for each test point x^* is defined as a multi-label classification task where each instance can be associated with some classes. The main advantage of this method is its flexibility because the selected experts depend on the given test point, and thus different experts can be assigned to different test points. Assume x^* is a new test point and $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ is the Gaussian experts set, and $\mathcal{L} = \{1, \ldots, M\}$ is the label set. The task is to find $\mathcal{M}^{\mathcal{C}}(x^*) = \{\mathcal{M}^{\mathcal{C}}{}_1(x^*), \ldots, \mathcal{M}^{\mathcal{C}}{}_K(x^*)\}$ that represents K selected experts to predict at x^*. We adopt two prominent classification models to solve this multi-label task without requiring problem transformations, K-nearest neighbors (KNN) and conventional deep neural networks (DNN).

KNN: Let $\mathcal{D}^{'} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$ be the partitions based on a disjoint partitioning strategy, i.e. K-Means clustering. Also, assume $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ contains the related centroids of the clusters in $\mathcal{D}$. For each test point x^*, there is a $1 \times M$ vector $dist(x^*, \mathcal{C})$, in which the i'th element is the distance between x^* and $\mathcal{D}_i$, where $dist()$ is a distance metric. The adopted KNN selects K experts with the closest centroids to x^*.

DNN: For each $x_i \in X$, $i = 1, \ldots, N$, the related partition label $l_i \in \mathcal{L}$ is available as an output of the training step in DGP. Therefore, instead of the original training set (X, y), a new set of points and labels $(X, \mathcal{L})$ is constructed for training the DNN. After that, for each test point, $x^* \in X^*$, the network will provide a probability vector $P^{\mathcal{L}}(x^*)$ where $P^{\mathcal{L}}(x^*)_j$ represents the probability that x^* belongs to the j'th expert. The K partitions with highest probabilities in $P^{\mathcal{L}}(x^*)$ are assigned to x^*.

4 Experiments

To evaluate the prediction quality of static and dynamic expert selection methods, we use *Concrete Compressive Strength* [2] data set. It is a small data set containing 1030 observations of 9 attributes (8 independent variables and one response variable). We use $\%90$ of the observations for training and the rest for testing, where disjoint partitioning is used to divide the data set into ten subsets. The quality of predictions is evaluated in two ways, standardized mean squared error (SMSE) and the mean standardized log loss (MSLL) [1]. Figure 1 depicts the prediction quality of expert selection methods on the *Concrete* data set for different values of K. The figure shows that multi-label-based expert selection models provide higher quality predictions with lower deviation from the NPAE model. The quality of the classification-based aggregations changes less as the selection parameter K increases. The case $K = 10$ leads to the original NPAE baseline (the dashed green lines), and both KNN and DNN return proper error values in both plots.



(a) SMSE (b) MSLL

Figure 1: **Expert Selection** prediction qualities of different experts selection methods compared to original baseline, NPAE, from *Concrete* data set.

5 Conclusion

In this work, we have proposed a novel expert selection approach for distributed learning with Gaussian agents, which leverages expert selection to aggregate dependent local experts' predictions. The available ensemble baselines use all correlated experts in the aggregation step. It affects the final predictions by the local predictions of weak experts or leads to impractically high computational costs. Our proposed approach uses a multi-label classification model, and the allocation of data points to experts is defined by considering the experts as class labels. Unlike the available expert selection method, the proposed model is more flexible to the model and data changes and chooses a related group of experts at each entry point. Excluding unrelated experts at each test point improves the prediction quality and reduces computational costs. The classification methods in this work, i.e., KNN and DNN, can be replaced with recent and more efficient solutions proposed to solve the multi-label classification problem. The proposed approach can be used for distributed and federated learning and does not impose restricted assumptions.

[2] `https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength`

References

[1] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[2] Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141, 2021.

[3] Aaron Zeng, Hodde Ho, and Yao Yu. Prediction of building electricity usage using gaussian process regression. *Journal of Building Engineering*, 28(11):101054, 2020.

[4] A. Denzel and J. Kästner. Gaussian process regression for geometry optimization. *he Journal of Chemical Physics*, 148(9):094114, 2018.

[5] R.B. Gramacy. *urrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC, 2020.

[6] T. Le, K. Nguyen, V. Nguyen, T. D. Nguyen, and D. Phung. Gogp: Fast online regression with gaussian processes. *IEEE International Conference on Data Mining*, pages 257–266, 2017.

[7] F. Tobar, T. D. Bui, and R. E. Turner. Learning stationary time series using gaussian processes with nonparametric kernels. *In Advances in Neural Information Processing Systems*, pages 3501–3509, 2015.

[8] M. P. Deisenroth and J. W. Ng. Distributed gaussian processes. *International Conference on Machine Learning*, pages 1481–1490, 2015.

[9] D. Peteiro-Barral and B. Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, 2013.

[10] J. Zhang, M. Wang, Q. Li, S. Wang, X. Chang, and B. Wang. A survey on distributed machine learning. *ACM Computing Surveys*, 53(2):1–33, 2020.

[11] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou. From distributed machine learning to federated learning: A survey. *arXiv preprint arXiv:2104.14362*, 2021.

[12] X. Shen H. Liu, Y. S. Ong and J. Cai. When gaussian process meets big data: A review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020.

[13] B. Szabó and H. van Zanten. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20(87):1–30, 2019.

[14] D. Rullière, N. Durrande, F. Bachoc, and C. Chevalier. Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28(4):849–867, 2018.

[15] H. Jalali, M. Pawelczyk, and G. Kasneci. Model selection in local approximation gaussian processes: A markov random fields approach. *IEEE International Conference on Big Data (Big Data)*, pages 768–778, 2021.

[16] H. Jalali, M. Pawelczyk, and G. Kasneci. Gaussian experts selection using graphical models. *arXiv preprint arXiv:2102.01496*, 2021.

[17] Francisco Herrera, Francisco Charte, Antonio J. Rivera, and María J. del Jesus. *Multilabel Classification*, pages 17–31. Springer International Publishing, 2016.

[18] C. Park and D. Apley. Patchwork kriging for large-scale gaussian process regression. *Journal of Machine Learning Research*, 19(7):1–43, 2018.

[19] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[20] Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

[21] V. Tresp. A bayesian committee machine. *Neural Computation*, 12(11):2719–2741, 2000.

[22] H. Liu, J. Cai, Y. Ong, and Y. Wang. Generalized robust bayesian committee machine for large-scale gaussian process regression. *International Conference on Machine Learning*, pages 1–10, 2018.

[23] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger. Unsupervised ensemble learning with dependent classifiers. *In Artificial Intelligence and Statistics*, pages 351–360, 2016.