Bayesian Sparse Regression for Mixed Multi-Responses with Application to Runtime Metrics Prediction in Fog Manufacturing

Xiaoyu Chen[1], Xiaoning Kang[2], Ran Jin[3], Xinwei Deng[4]

[1] Department of Industrial Engineering, University of Louisville, USA
[2] International Business College and Institute of Supply Chain Analytics, Dongbei University of Finance and Economics, China
[3] Grado Department of Industrial and Systems Engineering, Virginia Tech, USA
[4] Department of Statistics, Virginia Tech, USA

Fog computing (also referred as Edge computing) techniques have served as an important role in Industrial Internet of things (IIoT) for smart manufacturing systems. It provides local and distributed computation capabilities. The concept of Fog manufacturing is defined on integrating a Fog computing network with interconnected manufacturing processes, facilitates, and systems (See Figure 1). With local computation units (i.e., Fog units) close to the manufacturing processes, the Cloud-based centralized computation architecture can be evolved to a Cloud-Fog collaborative computation to provide higher responsiveness and significantly lower time latency. There is a trade-off between the local computing efficiency on a Fog unit and the global collaborative efficiency of the centralized Cloud. Specifically, the speciality of Fog units can significantly speedup the local computations, but it can pose significant challenges for the Cloud to assign the computation tasks and supervise the heterogeneous Fog units. Besides, fluctuated computation capability of the Fog units and intermittent communication conditions among the Fog units and the Cloud make it even harder for the collaboration. Therefore, computation offloading methods have been widely investigated to enable efficient collaboration between the Fog units and the Cloud with the consideration of constraints on resources.

In Fog manufacturing, the runtime performance metrics are often multivariate with mixed types. These metrics include the CPU utilization (i.e., continuous response), temperature of the CPU (i.e., continuous response), the number of computation tasks executed within a certain time period (i.e., counting response), and whether the memory utilization exceeds certain thresholds (i.e., binary response). Prediction and uncertainty quantification of these metrics are essential to support the computation in the Fog manufacturing, advancing analytics and optimization for high responsiveness and reliability. Based on the runtime performance metrics of these Fog nodes, the Fog computing can dynamically assign computation tasks to different Fog nodes. The manufacturing must provide responsive and reliable computation services by meeting all requirements in runtime performance metrics. It is thus of great importance to accurately predict runtime performance metrics of Fog nodes and quantify the uncertainty of

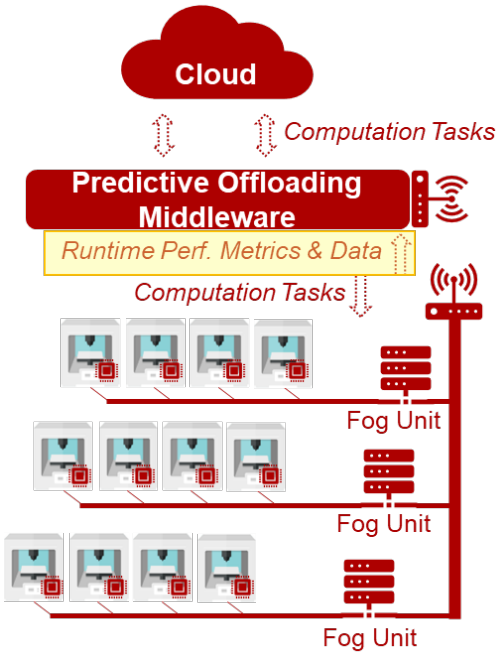

Figure 1. A Fog manufacturing framework with predictive offloading middleware to support reliable and responsive computation and communication service.

prediction in task assignment and offloading problems.

As the runtime performance metrics are multivariate with mixed types, a simple method is to model each individual metric separately. Clearly, such an approach overlooks the dependency relationship among the metrics, resulting in inaccurate prediction associated with high uncertainty. For example, as the increment in the executed number of computation tasks per minute (i.e., counting response), the CPU utilization and temperature (i.e., continuous responses) will increase. Quantifying such dependency among mixed responses is expected to improve the prediction accuracy. Moreover, by only providing point estimation of mixed responses, the model prediction may not be trustworthy for those with high prediction variance. Therefore, it calls for a joint model for mixed responses with uncertainty quantification. Towards predictive offloading, the objective is to jointly fit the mixed runtime performance metrics with the capability of statistical inferences to quantify uncertainties of the predicted metrics in Fog manufacturing.

In this work, we propose a Bayesian sparse multivariate regression for mixed responses (BS-MRMR) to achieve accurate model prediction and, more importantly, to obtain proper statistical inferences of the responses. The use of Bayesian estimation naturally enables uncertainty quantification of model prediction. Both group sparsity and individual sparsity are imposed on regression coefficients via proper spike-and-slab priors. The group structures often occur in the runtime performance metrics prediction problem when the metrics at the next time instance are regressed on two groups of predictors: the features extracted from the current and previous metrics (i.e., Group 1) and the covariates of the computation tasks (i.e., Group 2). On the other hand, not all predictors are important within each group. Hence individual sparsity is also induced for better estimation of model coefficients. Moreover, the proposed method considers the conditional dependency among multiple responses by a graphical model using the precision matrix, where a spike-and-slab prior is used to enable the sparse estimation of the graph. A Gibbs sampling scheme is then developed to efficiently conduct model estimation and inferences for the proposed BS-MRMR method. The proposed BS-MRMR model not only achieves accurate prediction, but also makes the predictive model more interpretable in the Fog manufacturing.

There are several directions for future studies. One direction is to investigate how to incorporate the quantified predicted uncertainty in the predictive offloading method by formulating the offloading problem as a chance-constrained optimization problem. Then the optimized offloading decisions can be more trustworthy to the performance of the predictive models. Besides predictive offloading, the proposed BS-MRMR model also facilitates the optimization of the Fog computing architecture by evaluating different designs based on the predicted performance metrics. Another direction is to extend the proposed BS-MRMR model to other types of responses such as censored outcomes and functional responses.

Reference

Chen, X., Kang, X., Jin, R., & Deng, X. (2023). Bayesian Sparse Regression for Mixed Multi-Responses with Application to Runtime Metrics Prediction in Fog Manufacturing. *Technometrics*, *65*(2), 206-219.