Bayesian inference for transfer learning

Loïc Iapteff, Julien Jacques, Matthieu Rolland, Benoit Celse

Bayesian inference for transfer learning

Loïc Iapteff [1], Julien Jacques [2], Matthieu Rolland [3] & Benoit Celse [4]

[1] *loic.iapteff @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[2] *julien.jacques @univ-lyon2.fr, 5 Avenue Pierre Mendès France, 69500 Bron*
[3] *matthieu.rolland @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[4] *benoit.celse @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*

Résumé. Le groupe IFP commercialise des catalyseurs et doit s'engager sur leur performance. Il est donc nécessaire de disposer de modèles prédictifs fiables pour chaque nouvelle génération de catalyseurs. Ces modèles sont construits à partir de données expérimentales très couteuses. Afin d'optimiser les coûts, notre ambition est de réduire le nombre d'expérimentations nécessaires pour estimer un modèle associé à un nouveau type de catalyseur, en transférant l'information contenue dans les modèles d'anciennes générations. Cet article décrit nos travaux sur le transfert de modèle linéaire par inférence bayésienne.

Mots-clés. Transfer learning, inférence bayésienne, modèle linéaire

Abstract. IFP group develops catalysts and has to guarantee their performances. It is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data. In order to minimize costs, we aim at reducing the number of new data points to measure to fit a model on the new catalyst, that is by using the knowledge available in the previous model. This paper describes our work on linear model transfer using Bayesian inference.

Keywords. Transfer learning, Bayesian inference, linear model

1 The problem

IFP group develops and sells catalysts to the chemical, bio chemical producers. Catalysts are solids that render the reaction feasible, faster, and / or at lower temperature and pressure. The performance must be guaranteed and it is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data and generally without accounting for the former catalysts' datasets. In order to minimize costs, we aim at reducing the number of new data points to measure on the new catalyst by transferring the models. By transferring, we mean use the knowledge available in the previous model and mix it with the new data points to build a good prediction model with a minimum of experimentation.

The problem is a transfer learning problem. Let's define a domain as $D = (X, P(X))$ with X a feature space and $P(X)$ its probability distribution, and an associated task

$T = (Y, f)$ with f the function used to predict $y \in Y$ given $\mathbf{x} \in X$. Pan & Yang (2010) define transfer learning as follow: "Given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t, transfer learning aims to help improve the learning of the target predictive function f_t in D_t using the knowledge in D_s and T_s, where $D_s \neq D_t$, or $T_S \neq T_T$". In our case, $D_s = D_t$ and $T_s \neq T_t$ as the catalyst and its performance is different, which put us in the inductive transfer learning case (Pan & Yang (2010)). The catalyst changes, so reaction will change and features have no reason to follow a particularly different distribution. There are different methods to solve these problems, such as transferring knowledge of instances, features or parameters.

We want a model to predict the output property Y_i with some information on feed and operating conditions, described using 12 features identical for the source and the target catalyst. Different models are tested to predict the output property on source data, specifically linear model, support vector, multi-layer perceptron, random forest, gradient boosting and kriging (Matheron (1969)). Best predictions are achieved with kriging but the linear model also offers satisfying results. Therefore, in this work the linear model is considered for its simplicity.

The model for the source catalyst is

$$Y_i = \beta_{s0} + \sum_{j=1}^{p} \beta_{sj} X_{ij} + \epsilon_i \tag{1}$$

with $\epsilon_i \sim \mathcal{N}(0, {\sigma_s}^2)$ and $p = 12$. Let $\hat{\theta}_s$ be the maximum likelihood estimate of $\theta_s = (\beta_{s0}, \ldots, \beta_{sp}, \sigma_s^2)$. The training data set available for the source catalyst is assumed to be sufficiently large such that $\hat{\theta}_s$ can be considered as satisfying estimate of θ_s.

Our goal is to estimate the same model but for the target catalyst

$$Y_i = \beta_{t0} + \sum_{j=1}^{p} \beta_{tj} X_{ij} + \epsilon_i \tag{2}$$

for which the available training data set is of a smaller size n_t.

We choose to focus on transfer knowledge of parameters because it's well adapted for the transformation of the linear model. Two approaches are considered. The first one is inspired from Bouveyron & Jacques (2010) and consists in identifying a link between θ_s and θ_t. The second one is a Bayesian approach and consider a Bayesian linear model for (2) with prior distribution on θ_t depending on θ_s.

2 Transfer models

The objective is to have a good model with as few points as possible. In the experiments presented in this paper, the performance of the transfer techniques are evaluated for

different numbers n_t of new points. The evaluation of the transferred model may depend on the sampling of training set. For this reason we present average results for 10 random samplings. For each sampling, RMSE score are evaluated on a test set independent of the training set. In the present industrial context, a model is considered satisfying if the RMSE score is lower than 0.005. With the source model the RMSE score is 0.0033 on the source data.

2.1 Transfer learning using parametric link

This method uses the idea of Bouveyron & Jacques (2010): keep some parameters unchanged for the target model ($\beta_{sj} = \beta_{tj}$ for some j), considering the influence doesn't change between both models, and then learn only others parameters.

If $\mathcal{M}$ is the set of index of parameters to be modified, then $\beta_{tj} = \beta_{sj}$ for $j \in \{1, ..., p\} \backslash \mathcal{M}$ and $\beta_{tj} = \lambda_j \beta_{sj}$ for $j \in M$. Only a reduced number of parameters have to be estimated for the target model. The challenge with this approach is the choice of $\mathcal{M}$. In this work $\mathcal{M}$ is selected by leave-one-out cross validation on the n_t target points, for all possible size of $\mathcal{M}$. With this approach, a performing model can be fitted with less points than if a totally new model is learned (Figure 1 left). For example, with 10 observations and modifying 1 parameter, RMSE is smaller than the objective of 0.005. In contrast, 30 observations are needed to a learned from scratch model to achieve such good results. With 100 observations, learned from scratch model is better. Changing a small number of parameters is more efficient for small training set but worse when a lot of data is available. An emerging challenge is the choice of the size of $\mathcal{M}$. Choosing a small size for $\mathcal{M}$ offers quick results. But by also trying to determine its size by cross validation, the results deteriorate.

Another remark, for a given size of $\mathcal{M}$, the modified parameters are not the same for small and large values of n_t where they do not change. In other words, we are not able to find the best parameters to transform with a few points. Assuming to know the best parameters to change, results are better in terms of number of target points, n_t (Figure 1 right). Parameters chosen to be modified are those chosen with 100 observations. So far, we have not been able to identify a technique to decide which are the best parameters to change on the available n_t points. For this reason, we explored Bayesian inference.

2.2 Transfer learning using Bayesian approach

In this section a Bayesian approach is used to learn parameters for the new linear model. The idea is to choose prior distributions depending on the source data. The model is then:

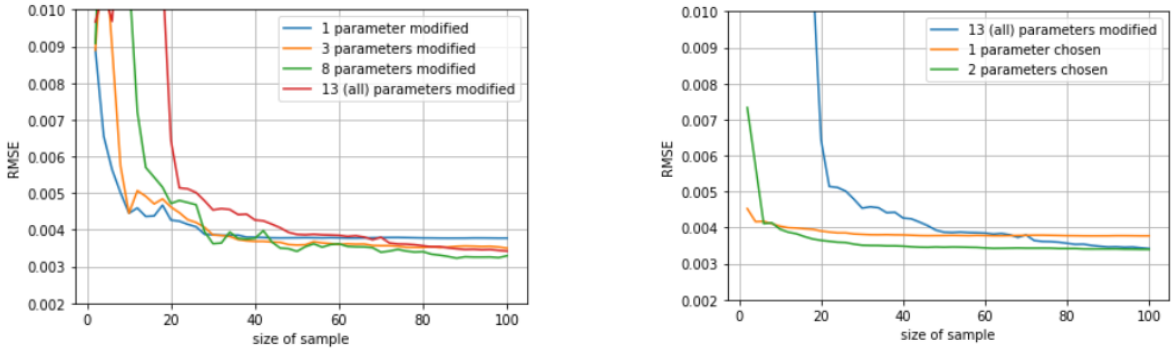

Figure 1: Graphs shows the evolution of RMSE according to n_t. On the left, parameters to modify are chosen by cross validation, on the right they are chosen knowing they are the best to modify.

$$\begin{aligned} Y_i &= \beta_{t0} + \sum_{j=1}^{p} \beta_{tj} X_{ij} + \epsilon_i, \\ \boldsymbol{\beta}_t &\sim \pi(\boldsymbol{\beta}_t), \\ \epsilon_i &\sim \mathcal{N}(0, {\sigma_t}^2), \\ \sigma_t &= \sigma_s, \end{aligned}$$

where $\boldsymbol{\beta}_t = (\beta_{t0}, \beta_{t1}, \ldots, \beta_{tp})^T$.

The Bayes Theorem gives that the posterior of $\boldsymbol{\beta}_t$ is

$$\pi(\boldsymbol{\beta}_t | \boldsymbol{Y}_t) = \frac{\pi(\boldsymbol{\beta}_t) f(\boldsymbol{Y}_t | \boldsymbol{\beta}_t)}{f(\boldsymbol{Y}_t)},$$

with $\boldsymbol{Y}_t = (Y_1, ..., Y_{n^t})^T$.

We consider different prior distribution $\pi(\boldsymbol{\beta}_t)$. The first one is the well known Zellner's prior (Zellner, 1986), also known as g-prior, for parameters $\boldsymbol{\beta}_t$:

$$\pi(\boldsymbol{\beta}_t) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}_s, g\sigma_t^2 (\boldsymbol{X}_t^T \boldsymbol{X}_t)^{-1}),$$

with $\widehat{\boldsymbol{\beta}}_s$ the maximum likelihood estimator (MLE) learned on the source data and $\boldsymbol{X}_t$ the $n^t \times (p+1)$ matrix of target observations, $(\boldsymbol{X}_t)_{i1} = 1$ for $i = 1, ..., n$.

Using such a prior, only the mean of the prior distribution depends on the source data. The variance of the prior depends only on target data and on a fixed parameter g. Notice that the posterior's mean using such a prior is a weighted average between the MLE and the mean of the prior. In our case, a weighted average between the MLE for source data and the MLE for target data is calculated: $\widehat{\boldsymbol{\beta}}_t = \frac{1}{g+1}(g\widehat{\boldsymbol{\beta}}_{MLE,t} + \widehat{\boldsymbol{\beta}}_s)$.

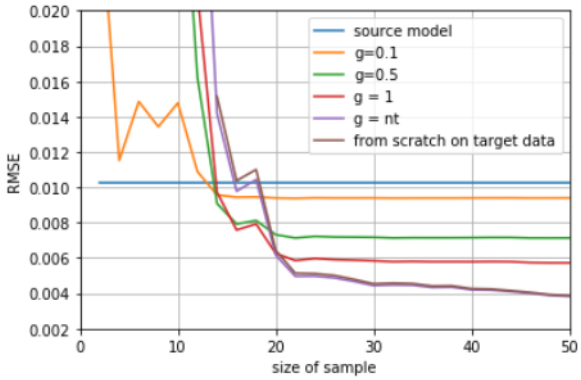

Figure 2: Comparison between an estimation of $\boldsymbol{\beta}_t$ with a Bayesian approach and a g-prior for different values of g, and a model learned without any prior.

With this prior, results are not satisfying whatever the value of g. $\widehat{\boldsymbol{\beta}}_s$ is not a good estimator for $\boldsymbol{\beta}_t$, thus averaging with the MLE for target doesn't improve results.

An idea to improve the results is to increase the information transferred. Source data is employed to learn both the mean and the variance of prior. Once again, a Gaussian prior is considered for parameters $\boldsymbol{\beta}_t$. The prior mean is $\widehat{\boldsymbol{\beta}}_s$. The variance is chosen as the variance of $\widehat{\boldsymbol{\beta}}_s$ scaled with a scalar λ. Let remark that when $\lambda = 1$, this prior for $\boldsymbol{\beta}_t$ corresponds to the posterior distribution of $\boldsymbol{\beta}_s$ estimated in a Bayesian model with an uniform prior for $\boldsymbol{\beta}_s$. The introduction of the factor λ allows more flat prior:

$$\pi(\boldsymbol{\beta}_t) = \mathcal{N}(\widehat{\boldsymbol{\beta}}_s, \lambda\sigma_s^2(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}).$$

The mean of the corresponding posterior distribution leads to:

$$\widehat{\boldsymbol{\beta}}_t = (\boldsymbol{X}_t^T\boldsymbol{X}_t + \sigma_t^2\lambda^{-1}\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{X}_t^T\boldsymbol{Y}_t + \sigma_t^2\lambda^{-1}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\beta}}_s),$$

with $\boldsymbol{\Sigma} = \sigma_s^2(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}$.

The parameter value λ must yield the best performance with the fewest observations possible. When $\lambda \to \infty$ the posterior mean tends to the MLE. When $\lambda \to 0$ the posterior mean tends to the prior. To see the impact of lambda, different values are tried and RMSE evolution is evaluated (Figure 3 left). There is an optimum in the λ value in the range $100 - 1000$ for our data.

Data are normalized, thus parameters $\boldsymbol{\beta_t}$ take values in $[-1, 1]$. A compromise must be found between prior information and variability of parameters. A good choice of standard deviation for each parameter seems to be near from 1 to cover the $[-1, 1]$ interval and not be too wide. Taking a $\lambda = (mean((\Sigma_{jj})_{j=0,..,p}))^{-1} \simeq 800$ on our data yields an average variance of 1 for each parameter. This value of λ offers good results with few points (Figure 3 right). With this approach, performing models can be fitted with a number of target points smaller than without knowledge on old catalyst. RMSE score of 0.004 can be reached with only 5 target points instead of 50 for a learned from scratch model.

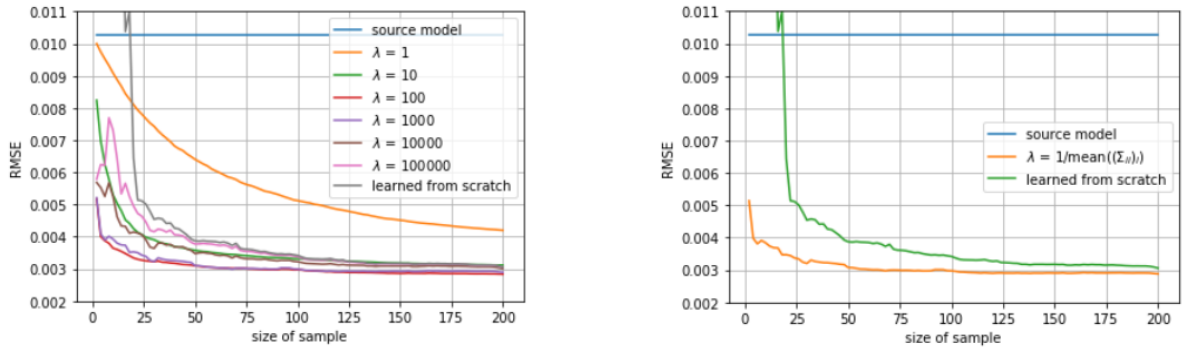

Figure 3: Impact of λ on $\pi(\boldsymbol{\beta}_t|\boldsymbol{Y}_t)$.

3 Perspectives

Bayesian inference with an optimized learning parameter (lambda) gives good results on transfering linear models on our data set. The knowledge from an old catalyst, namely the parameter covariance matrix and values, is shown to improve the quality of the linear model for the new catalyst. Future works will focus on 2 mains areas: the transfer of kriging models, still with a bayesian transfer knowledge, and the design of experiments by choosing the best target data to use for transfer.

References

Bouveyron, C., & Jacques, J. (2010). Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters*, 2237-2247.

Matheron, G. (1969) Le krigeage universel (Universal kriging) Vol. 1. Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau, 83 p.

Pan, S., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1345-1359.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P. and Zellner, A., Eds., Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Elsevier Science Publishers, Inc., New York, 233-243