Probabilistic Expert Knowledge Elicitation of Feature Relevances in Sparse Linear Regression

Pedram Daee*, Tomi Peltola* Marta Soare*, and Samuel Kaski

Helsinki Institute for Information Technology HIIT and
Department of Computer Science, Aalto University, Finland,
firstname.lastname@aalto.fi
*Authors contributed equally.

1 Introduction

In this extended abstract[1], we consider the "small n, large p" prediction problem, where the number of available samples n is much smaller compared to the number of covariates p. This challenging setting is common for multiple applications, such as precision medicine, where obtaining additional samples can be extremely costly or even impossible. Extensive research effort has recently been dedicated to finding principled solutions for accurate prediction. However, a valuable source of additional information, domain experts, has not yet been efficiently exploited.

We propose to integrate expert knowledge as an additional source of information in high-dimensional sparse linear regression. We assume that the expert has knowledge on the relevance of the features in the regression and formulate the knowledge elicitation as a sequential probabilistic inference process with the aim of improving predictions. We introduce a strategy that uses Bayesian experimental design [2] to sequentially identify the most informative features on which to query the expert knowledge. By interactively eliciting and incorporating expert knowledge, our approach fits into the interactive learning literature [1, 8]. The ultimate goal is to make the interaction as effortless as possible for the expert. This is achieved by identifying the most informative features on which to query expert feedback and asking about them first.

2 Method

We introduce a probabilistic model that subsumes both a sparse regression model which predicts external targets, and a model for encoding expert knowledge. We then present a method to query expert knowledge sequentially (one feature at a time), with the aim of getting fast improvement in the predictive accuracy of the regression with a small number of queries.

For the regression, a Gaussian observation model with a spike-and-slab sparsity-inducing prior [5] on the regression coefficients is used: $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I})$, $w_j \sim \gamma_j \, \mathrm{N}(0, \psi^2) + (1 - \gamma_j)\delta_0; \gamma_j \sim \mathrm{Bernoulli}(\rho), j = 1, \ldots, p$, where $\boldsymbol{y} \in \mathbb{R}^n$ are

[1] This extended abstract is adapted from [3].

the output values and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the matrix of covariate values. The regression coefficients are denoted by $w_1, \ldots, w_p$, and σ^2 is the residual variance. The γ_j indicate inclusion ($\gamma_j = 1$) or exclusion ($\gamma_j = 0$) of the covariates in the regression (δ_0 is a point mass at zero). The prior expected sparsity is controlled by ρ. The expert knowledge on the relevance of the features for the regression is encoded by a feedback model: $f_j \sim \gamma_j \operatorname{Bernoulli}(\pi) + (1 - \gamma_j) \operatorname{Bernoulli}(1 - \pi)$, where $f_j = 1$ indicates that feature j is *relevant* and $f_j = 0$ *not-relevant*, and π is the probability that the expert feedback is correct relative to the state of the covariate inclusion indicator γ_j.

As the number of covariates p can be large, we assume that it is infeasible, or at least unnecessarily burdensome, to ask the expert about each feature. Instead, we aim to ask first about the features that are estimated to be the most informative given the (small) training data, and frame this problem as a Bayesian experimental design task [2, 9]. We prioritize features based on their expected information gain for the predictive distribution of the regression. As the expert is queried for the feedbacks sequentially, the posterior distribution of the model and the prioritization are recomputed after each feedback in order to use the latest knowledge. At iteration t for feature j, the expected information gain is

$$\mathrm{E}_{p(\tilde{f}_j|\mathcal{D}_t)} \left[\sum_i \mathrm{KL}[p(\tilde{y}|\mathcal{D}_t, \boldsymbol{x}_i, \tilde{f}_j) \parallel p(\tilde{y}|\mathcal{D}_t, \boldsymbol{x}_i)] \right],$$

where $\mathcal{D}_t = \{(y_i, x_i) : i = 1, \ldots, n\} \cup \{f_{j_1}, \ldots, f_{j_{t-1}}\}$ denotes the training data together with the feedback that has been given at previous iterations and $p(\tilde{f}_j|\mathcal{D}_t)$ is the posterior predictive distribution of the feedback for the jth feature. The summation over i goes over the training dataset. This query scheme goes beyond pure prior elicitation [4, 6, 7] as the training data is used to facilitate an efficient expert knowledge elicitation. This is a crucial aspect that enables the elicitation in high-dimensional regression.

3 Discussion

The proposed method was tested in several "small n,large p" scenarios on synthetic and real data with simulated and real users [3]. The results confirm that improved prediction accuracy is already possible with a small number of user interactions, for the task of predicting product ratings based on the relevance of some of the words used in textual reviews. Our method can naturally be used on many other applications where expert feedback is needed, its main advantage being that it efficiently reduces the burden on the expert by asking first the most informative queries. However, the amount of improvement in different applications depends on the type of feedback requested, and on willingness and confidence of experts to provide the feedback. In addition, appropriate interface and visualization techniques are also required for a complete and effective interactive elicitation. These considerations are left for future work.

Acknowledgements This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN; grants 295503, 294238, 292334, and 284642), Re:Know funded by TEKES, and MindSee (FP7–ICT; Grant Agreement no 611570).

References

1. Amershi, S.: Designing for Effective End-User Interaction with Machine Learning. Ph.D. thesis, University of Washington (2012)
2. Chaloner, K., Verdinelli, I.: Bayesian experimental design: A review. Statistical Science 10(3), 273–304 (08 1995)
3. Daee, P., Peltola, T., Soare, M., Kaski, S.: Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. Machine Learning (Jul 2017), `https://doi.org/10.1007/s10994-017-5651-7`
4. Garthwaite, P.H., Dickey, J.M.: Quantifying expert opinion in linear regression problems. Journal of the Royal Statistical Society. Series B (Methodological) pp. 462–474 (1988)
5. George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. Journal of the American Statistical Association 88(423), 881–889 (1993)
6. Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., Peters, S.C.: Interactive elicitation of opinion for a normal linear model. Journal of the American Statistical Association 75(372), 845–854 (1980)
7. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain Judgements. Eliciting Experts' Probabilisties. Wiley, Chichester, England (2006)
8. Porter, R., Theiler, J., Hush, D.: Interactive machine learning in data exploitation. Computing in Science & Engineering 15(5), 12–20 (2013)
9. Seeger, M.W.: Bayesian inference and optimal design for the sparse linear model. Journal of Machine Learning Research 9, 759–813 (2008)