A Novel Hybrid Active Learning Strategy for Nonlinear Regression

Karel Crombecq
University of Antwerp
Antwerp, Belgium
Karel.Crombecq@ua.ac.be

Ivo Couckuyt
Ghent University
Ghent, Belgium

Eric Laermans
Ghent University
Ghent, Belgium

Tom Dhaene
Ghent University
Ghent, Belgium

Abstract

In many supervised machine learning problems, the labeling of data points is a very expensive operation, often requiring the intervention of a human agent to accomplish the task. Therefore, many methods have been developped to minimize the number of labeled data points required to achieve a certain accuracy. In active learning, information gathered from previously labeled data is used to determine which data points need to be labeled next. This guarantees more efficient resource usage by focusing data points on areas which are estimated to be interesting or which have a high uncertainty. In this paper, we propose a novel hybrid exploration-exploitation active learning strategy.

1 Introduction

In regression learning, the goal is to find an accurate approximation of a continuous output as a function of a set of (potentially) continuous features. In traditional supervised learning, the output value for each data point is known in advance, and the model is trained using all the data that is available. However, this may be infeasible for several reasons. Firstly, the dataset might be so large that training the model on the entire dataset is too expensive. Secondly, it may be too expensive to acquire the output value for each data point. For example, the output value may be generated from a high fidelity computer simulation, which can take hours or even days for one single evaluation (Gorissen et al., 2008). Finding the output value may also imply a physical experiment, which requires interaction with a human agent to set up.

Active learning attempts to address these problems by iteratively selecting new data points from the dataset for evaluation. This is illustrated in Figure 1. During each iteration, a model or a set of models is trained on the available (evaluated) data points, and a decision is made whether or not the algorithm should select new data points to achieve the desired accuracy. This makes it possible to use information from previous iterations (model accuracy, model uncertainty, previously acquired outputs) to select new data points as optimally as possible. For example, Farhang-Mehr and Azarm (2005) propose an algorithm that uses previously built models to find regions in which many local optima of the surrogate model are close to each other, and selects new samples in these regions.

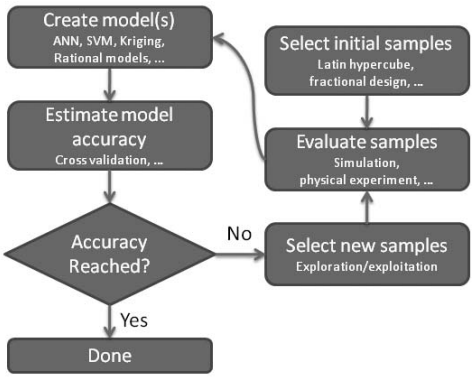

Figure 1: A flow chart of a typical active learning strategy.

2 Hybrid active learning

An essential consideration in active learning is the trade-off between exploration (input-based sampling) and exploitation (output-based sampling). Exploration is the act of filling up the design space as evenly as possible in order to identify key regions of the design space, such as discontinuities, steep slopes, optima, stable regions and so on, depending on the objective. Exploration does not involve the responses of the system, because the goal is to fill up the input domain evenly.

Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning, pages 110–111,
Tilburg, May 2009.

Exploitation is the alternative. Instead of exploring the domain, data points are selected in regions which have already been identified as (potentially) interesting. For example, one might want to zoom in on optima, in order to make sure the regression model does not overshoot the optimum. Or one might also want to investigate near possible discontinuities to verify that they are, in fact, discontinuous, and not just very steep slopes. Finally, one might want to sample in regions in which the current best model is very uncertain. Exploitation involves using the outputs of the previous data points and previously trained regression models to guide the sampling process.

In every active learning strategy, a trade-off must be made between these two conflicting options. If an active learning strategy only focuses on exploitation, large regions of the design space may remain unsampled, and interesting regions may be overlooked completely. On the other hand, if a strategy only focuses on exploration, the advantage provided by evaluating and selecting the samples iteratively is ignored, because the outputs are not used. This means that any competitive sequential design strategy must somehow accomodate these two options.

We propose a new hybrid active learning method which performs a trade-off between exploration and exploitation. New samples will be selected in regions which are either highly non-linear (exploitation) or undersampled (exploration). Exploration is achieved by approximating a Voronoi diagram of the previously evaluated data points. The Voronoi cell size for each point is then used as an estimation of the sampling density in that region. New samples are selected in the largest Voronoi cells. Exploitation is achieved by estimating the gradient of the function at each data point, because the gradient is the best local linear approximation (LOLA) of the model. Next, the deviation is calculated between the local linear approximation and the neighbouring data points. If there is a high deviation, the region around the data point must be non-linear, and thus new samples will be chosen nearby.

3 Example

We consider the following simple one-dimensional academic test function:

$$f(x) = 3(x+1) + e^{3.5x} sin(8\pi x) \qquad (1)$$

This function is shown in 2, along with the samples that were adaptively chosen by the active learning strategy outlined above. On the left side, the function behaves almost linearly, while on the right side the function oscillates wildly. The plot clearly shows that the hybrid active learning strategy selected samples more densely in highly nonlinear regions of the design space, demonstrating that the method is capable of quickly identifying nonlinear regions without neglecting linear regions completely.

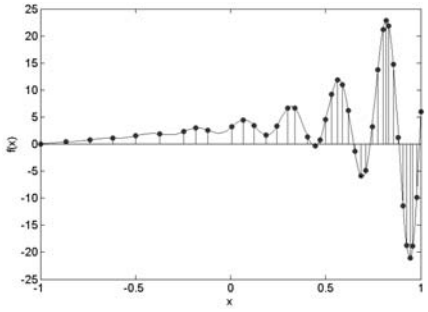

Figure 2: The function described in Equation 1. The data points that were adaptively selected are displayed as dots.

4 Conclusion

The new hybrid active learning strategy proposed in this paper performs an effective trade-off between exploration and exploitation, combining the advantages of each method to produce a robust, scalable and flexible active learning strategy. For now, this method can only be applied to continuous (real or complex) outputs. Future work will include investigating the trade-off between exploration and exploitation and using dynamic weighting systems to balance them more efficiently. The method will also be adapted to work with discrete outputs, so that it can be used to select data points in classification tasks.

References

A. Farhang-Mehr and S. Azarm. 2005. Bayesian meta-modelling of engineering design simulations: a sequential approach with adaptation to irregularities in the response behaviour. *International Journal for Numerical Methods in Engineering*, 62(15).

D. Gorissen, K. Crombecq, I. Couckuyt, and T. Dhaene. 2008. Automatic approximation of expensive functions with active learning. In *Foundation on Computational Intelligence, Learning and Approximation*. Springer Verlag.