Online Parameter Selection for Gas Distribution Mapping

V. Hernandez Bennetts†, M. Trincavelli, E. Schaffernicht, A.J. Lilienthal
Centre for Applied Autonomous Sensor Systems, Örebro University, Örebro, Sweden
†E-mail: victor.hernandez@oru.se

Keywords: Virtual leave-one-out cross validation, model selection, gas distribution mapping

ABSTRACT

Kernel DM+V, one of the most common and well studied algorithms for gas distribution mapping, relies crucially on a set of parameters. These parameters are often estimated using cross-validation (CV), which is computationally expensive and therefore has to be carried out offline. Here we propose an efficient method for parameter selection, based on Virtual Leave-One-Out Cross Validation (VLOOCV), which enables online calculation of the optimal set of parameters. VLOOCV approximates the results of CV at greatly reduced computational costs. We validate the proposed method in one indoor experiment where a mobile robot with a Photo Ionization Detector (PID) was collecting gas measurements while moving in the target area. A comparison with the standard measure for model selection, the CV–based NLPD (Negative Log Predictive Density), favors the proposed algorithm that achieves the same model selection performance using just a fraction of the computational resources.

ALGORITHM

Mapping the distribution of one or multiple gases using a mobile robot is a challenging task. Kernel DM+V [1], probably the most robust approach today, interprets the mapping problem as a kernel regression method. At the core of the algorithm is the well known Nadaraya-Watson estimator with RBF kernel [2], which is applied twice, once for the estimation of the predictive mean and once for the estimation of the predictive variance. Kernel regression depends crucially on the choice an appropriate bandwidth of the kernel. It is usually selected in a grid search over the kernel bandwidth space using the average NLPD obtained from CV as evaluation criterion. The computational cost of this method is high due to the need for building and evaluating multiple models on different parts of the data for cross-validation. Instead, VLOOCV builds a single model using the whole dataset and estimates the effect of leaving out parts of the data [3]. To this end, VLOOCV calculates a leverage score h_i for each data point i indicating the effect of the point on the model. Scores of zero correspond to points that do not affect the model. To estimate the NLPD using VLOOCV, the negative likelihood of each of the training points according to the model built with the whole training set is computed and then weighted by h_i according to the following formula:

$$NL_i = \frac{1}{2(1-h_i)}\left(\log v_i + \frac{r_i - m_i}{v_i}\right) \quad (1)$$

Where m_i and v_i are the mean and variance of the distribution given by the model built on the whole training set. The NLPD estimate is given by the average of all the NL_i scores. The computational complexity of the update of the CV and VLOOCV algorithms is dominated by the update of the variance map and the calculation of the NL_i and h_i scores which scale linearly with the number of points. The computational advantage of the VLOOCV is due to the fact that only 1 model instead of (number of CV-folds -1) need to be updated.

RESULTS

The proposed parameter selection algorithm has been tested on data collected with a PID

sensor mounted on a mobile robot deployed in one indoor (Fig.1) and one outdoor environment where a gas source releasing ethanol was placed. Fig. 2 shows good agreement for the value of the NLPD calculated using CV, VLOOCV and VLOOCV without h_i score correction. Fig. 3 reports the computational time of the three algorithms for a different number of measurement points. The VLOOCV algorithm is computationally least expensive, while preserving the properties of the objective function (clear minimum in correspondence of the selected bandwidth). It is worth noting that correcting the negative likelihood with leverage scores does hardly change the VLOOCV result. This means that computation of the leverage factors is not needed for gas distribution mapping data obtained with mobile robots. An explanation is that the models obtained with the full dataset and with a fraction of the dataset (used in CV) are very similar, as shown in Fig. 4. The score displayed is the intersection of the Normal distributions (1 identical distributions, 0 distributions that do not intersect) predicted by the model computed with the full dataset and only with one fold.

ACKNOWLEDGEMENT

The authors would like to thank Robotdalen for supporting the Gasbot project (project number 8140).

Reference

1. A. J. Lilienthal, M. Reggente, M. Trincavelli, J.L. Blanco and J. Gonzalez, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* **(2009)**, pp. 570-576.

2.Nadaraya, E. A. in *Theory of Probability and its Applications* **(1964)**, Vol. 9, Issue 1, pp. 141–2.

3. G. Monari and G. Dreyfus, in *Neural Computation*, MIT Press **(2002)**, Vol. 14, pp.1481-1506.

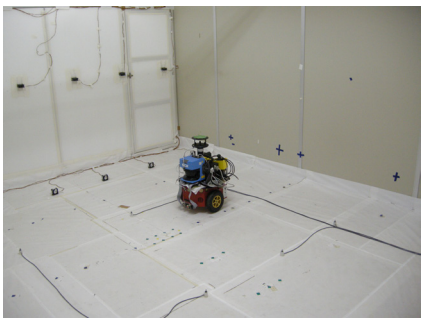

Figure 1: Robot during indoor experiment.

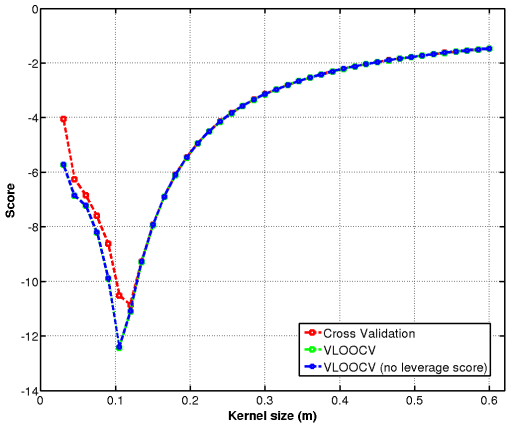

Figure 2: NLPD scores for the CV, VLOOCV and VLOOCV without correction scores. The scores obtained by VLOOCV and the VLOOCV without corrections are basically identical.

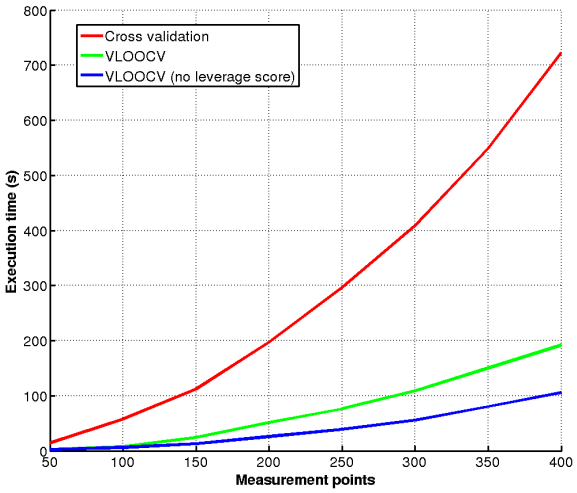

Figure 3: Computational time for the three algorithms. As expected the computation times scale roughly linearly with the number of measurements.

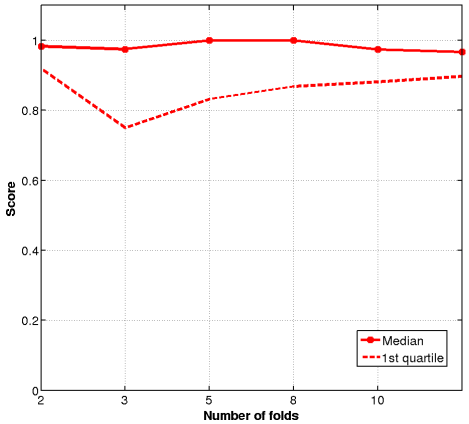

Figure 4: Similarity scores between the model calculated on the full dataset and the models calculated on fractions of the dataset (1/number of folds).