

Available online at <http://www.mecs-press.net/ijem>

Local Reweighted Kernel Regression

Weiwei Han

Department of Mathematics & Computer Science of Guangdong University of Business Studies, Guangzhou, 510320, China

Abstract

Estimating the irregular function with multiscale structure is a hard problem. The results achieved by the traditional kernel learning are often unsatisfactory, since underfitting and overfitting cannot be simultaneously avoided, and the performance relative to boundary is often unsatisfactory. In this paper, we investigate the data-based localized reweighted regression model under kernel trick and propose an iterative method to solve the kernel regression problem. The new framework of kernel learning approach includes two parts. First, an improved Nadaraya-Watson estimator based on blockwisely approach is constructed; second, an iterative kernel learning method is introduced in a series decreased active set to choose kernels. Experiments on simulated and real data sets demonstrate that the proposed method can avoid underfitting and overfitting simultaneously and improve the performance relative to the boundary effect.

Index Terms: irregular function; statistic learning; multiple kernel learning

© 2011 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science.

1. Introduction

Here introduce the paper, and put a nomenclature if necessary, in a box with the same font size as the rest of the paper. The paragraphs continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers, bold and 10 pt. Here follows further instructions for authors. Learning to fit data with noise is an important research problem in many real-world data mining applications. Kernel tricks have attracted more and more research attention recently. Fro given data set $S = \{(x_i, y_i)\}_{i=1}^n$, the solution of the kernel learning problem is of the form

$$m(x) = \sum_{i=1}^n \alpha_i K(x, x_i) \quad (1)$$

* Corresponding author.
E-mail address:

Where α_i are the coefficients to be learned from examples, while K is positive definite kernel associated with RKHS H . Recently, using multiple kernels instead of a single one can enhance the interpretability and improve performance [1]. In such cases, a convenient approach is to consider:

$$K(x, x') = \sum_{i=1}^N c_i K_i(x, x'), \text{ s.t. } \sum_{i=1}^N c_i = 1, c_i \geq 0 \quad (2)$$

Where N is the total number of kernels. In general, the correct number of kernels is unknown, and simultaneously determining the required number of kernels as well as estimating the associated parameters of MKL is a challenging problem [1].

For non-flat functions which comprise both the steep variations and the smooth variations, it is sometimes unsuitable to use one kernel even if a composite multiple kernel with several global bandwidths to estimate the unknown function [2]. First, the kernels are chosen prior to learning, which may be not adaptive to the characteristics of the function so that underfittings and overfittings occur frequently in the estimated function [3]. Although, the localized multiple kernel learning proposed in is adaptive to portions of high and low curvature, it is sensitive to initial parameters. Second, how to determine the number of kernels is unanswered. Finally, classical kernel regression methods exhibit a poor boundary performance [5] [6] [7].

The rest of this correspondence is organized as follows. In section 2, we proposed an iterative localized regression to deal with nonflat function regression problem. Section 3 presented regression results on numerical experiments on synthesis and real-world data sets while section 4 concludes the paper and contains remarks and other issues about future work.

2. Localized reweighted regression method

2.1. Related Work

Nadarya (1964) [8] and Watson (1964) [9] proposed to estimate $m(x)$ using a kernel as a weighting function. Given the sample data set $S = \{(x_i, y_i)\}_{i=1}^n$:

$$\hat{m}(x; S) = \frac{\sum_{i=1}^n K_h(x, x_i) y_i}{\sum_{i=1}^n K_h(x, x_i)} = \sum_{i=1}^n w_i(x; S) y_i$$

Where $w_i(x; S) = \left[\sum_{i=1}^n K_h(x, x_i) \right]^{-1} K_h(x, x_i)$ is the Nadaraya-Watson weights, such that

$$\sum_{j=1}^n w_j(x; S) = 1, \forall x$$

And $K_h(x) = h^{-1} K(x/h)$ is a kernel with bandwidth h .

Associating blockwise technique, we propose an improved localized kernel regression estimator which achieves automatic data-driven bandwidth selection. Suppose the initial data set S is partitioned into p

blocks denote by SS_1, SS_2, \dots, SS_p with length d_1, d_2, \dots, d_p such that $\sum_{j=1}^p b_j = n$ [11]. For given x , if

there is some block SS_x such that

$$\min\{x_i \mid x_i \in SS_x\} \leq x \leq \max\{x_i \mid x_i \in SS_x\}$$

Then the blockwised Nadaraya-Watson estimator is given as follows

$$\hat{m}(x; SS_x) = \frac{\sum_{x_i \in SS_x} K_h(x, x_i) y_i}{\sum_{x_i \in SS_x} K_h(x, x_i)} = \sum_{x_i \in SS_x} w_i(x; SS_x) y_i$$

As thus, the localized estimator presents the unknown function m without a complicated parameters selection procedure.

2.2. The new regression method

Given a dataset $S = \{(x_i, y_i), x_i \in R^n, y_i \in R\}$. Assume that $m(x) \in H$, where H is some reproducing kernel Hilbert space called active space, with respect to the reproducing kernel K . The square norm related to the inner product by $\|f\|_H^2 = \langle f, f \rangle_H$. Consider the problem,

$$\min H(m) = \sum_{i=1}^n L(y_i, m(x_i)) + \lambda P(m) \quad (3)$$

Where λ is a positive number which balances the trade-off between fitness and smoothness; L is a loss function; $P(m) = \|m\|_H^2$ is penalized function. By the represent theory, the solution of equation (3) is [12],

$$\hat{m}(x) = \sum_{i=1}^n \alpha_i K(x, x_i) \quad (4)$$

A generalized framework of kernel is defined as

$$K(x, x') = \sum_{i=1}^N c_i K_i(x, x') \quad (5)$$

Where $K_i, i=1, \dots, N$ are N positive definite kernels on the same input space X , and each of them being associated to a RKHS H_i whose elements will be denoted f_i and endowed with an inner product $\langle \cdot, \cdot \rangle_i$, and $\{c_i\}_{i=1}^N$ are coefficients to be learned under the nonnegative and unity constraints

$$\sum_{i=1}^N c_i = 1, c_i > 0, 1 \leq i \leq N \quad (6)$$

How to determine N is an unanswered problem. For any $c_i > 0$, H'_i is the Hilbert space derived from H_i as follows:

$$H'_i = \{f \mid f \in H_i : \frac{\|f\|_{H_i}}{c_i} < \infty\}$$

Endowed with the inner product

$$\langle f, g \rangle_{H'_i} = \frac{1}{c_i} \langle f, g \rangle_i$$

Within this framework, H'_i is a RKHS with kernel $K'_i = c_i K_i(x, x')$, since

$$m(x) = \langle m(\cdot), K_i(x, \cdot) \rangle_{H_i} = \langle m(\cdot), c_i K_i(x, \cdot) \rangle_{H_i'}$$

Then, we define H as the direct sum of the RKHS H_i' . Substituting (5) into (4), an updated equation of (2) is obtained as follows,

$$\begin{aligned} \hat{m}(x) &= \sum_{i=1}^n \alpha_i K(x, x_i) \\ &= \sum_{i=1}^n \alpha_i \sum_{j=1}^N c_j K_j(x, x_i) \\ &= \sum_{j=1}^N c_j \sum_{i=1}^n \alpha_i K_j(x, x_i) = \sum_{j=1}^N m_j(x) \end{aligned} \quad (7)$$

Instead of the equation (3), we convert to consider the models for $j = 1, \dots, N$,

$$\min H_j(m_j) = \sum_{i=1}^n L(y_i, m_j(x_i)) + \lambda_j P(m_j) \quad (8)$$

Then, the kernel learning problem can thus be envisioned as learning a predictor belonging to a series of adaptive hypothesis space endowed with a kernel function. The forthcoming part explains how we solve this problem. Assume that a kernel function $K_1(\cdot, \cdot)$ and corresponding reproducing kernel Hilbert space H_1' are included, and then we get the initial estimator,

$$\hat{m}_1(x) = \frac{\sum_{j=1}^p \sum_{x_i \in SS_j} K_h(x, x_i) y_i}{\sum_{j=1}^p \sum_{x_i \in SS_j} K_h(x, x_i)} \quad (9)$$

The residual function can be obtained,

$$res_1(x) = m(x) - \hat{m}_1(x) \in V = H - H_1' \quad (10)$$

If we have introduced t kernels $\{K_j\}_{j=1}^t$, then the estimator can be updated as

$$\hat{m}(x) = \sum_{j=1}^t \hat{m}_j(x) = \sum_{j=1}^t \sum_{i=1}^n \alpha_i^j K_j(x, x_i)$$

And the residual function,

$$res_t(x) = m(x) - \hat{m}(x) \quad (11)$$

If the measurements of res_t fulfilled certain thresholding criteria, here we employ 2-norm, $N = t$ represents the number of introduced kernels and puts an end to iteration procedure. If not, considering the problem in the decreased subspace H'_{t+1} , compute $res_t = y_i - \hat{m}(x_i)$ and update the sample set $S = \{(x_i, res_t)\}$ which can be treated as the limited of the initial data set in H'_{t+1} .

Employing iteration, we will consider a new regression problem on the updated sample data set in a decreased subspace.

Compared with the general MKL, the first advantage is that it needs not to select weights α_i which will reduce much more computation burden and just need to select one kernel bandwidth at each iteration step.

Furthermore, the new method is adaptive to the local curvature variation and improves the boundary performance as a result of the introduction of blockwised Nadaraya-Watson estimate technique. At last, the number of kernels introduced will change according to real data settings based on iteration which will avoid underfitting and overfitting problem effectively.

2.3. Algorithm

The complete algorithm of Iterative Localized Kernel Reweighted Regression can be briefly described by the following steps:

- 1) Input S , the maximum iteration step M , threshold ε , $N = 1$;
- 2) Initialize the pilot estimator $\hat{m}(x) = 0$, and pilot residual $e = y$;
- 3) Update the data set $S = \{(x_j, e_j)\}_{j=1}^n$;
- 4) Select kernel K , compute the estimator $\hat{m}(x; S)$ with equation (9);
- 5) Update the estimator $m(x) = m(x) + \hat{m}(x; S)$;
- 6) Update the residuals $e = y - m(x)$, and $N = N + 1$;
- 7) Calculate the norm of residual e .

Repeating the steps from 3) to 7), this process is continued until the norm of residual e is smaller than the pre-determined value ε or the iteration step N is larger than the pre-determined value M .

In order to select parameters, we choose 10-fold cross-validation: randomly divide the given data into ten blocks and consider the Generalized Cross Validation function is given as

$$GCV(\theta) = \frac{1}{10} \sum_j (\hat{m}^{(-j)} - m)^2$$

Where θ represents the set of relevant parameters, and $\hat{m}^{(-j)}$ is the estimator of m without the j th block samples of S .

3. Experiment results

We have conducted studies on simulated data and real-world data using the proposed method.

3.1. Application to Simulated Data

The test function is the mixture of Gaussian and Laplacian distributions define by

$$m(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} + \frac{0.7}{4} e^{-0.7|x+2|}$$

The number of data points for experiment is 200, and the experiment was repeated 50 times. Figure 1(a) shows the target values which were corrupted by white noise. The performance of the experiment was shown in Figure 1, in which the slender line present the true test function and the bold line represented the estimated results. Figure 1(b) represented the estimated curve using the proposed method with two step iteration which deals well with different portions with different curvature; Figure 1(c) demonstrated the standard single kernel regression based on Gaussian kernel with a global bandwidth. For this example, it can be seen that the Iterative Localized Kernel regression method achieved the better performance. Compared with the proposed method, the single kernel regression could not avoid underfitting and overfitting simultaneously and sensitive to noise at the boundary area.

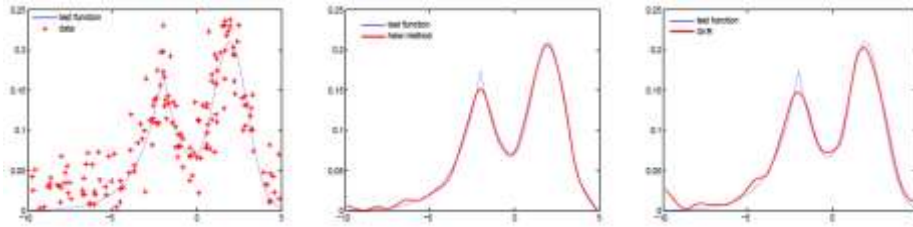


Fig. 1. The test function (slender line) and the approximation function (bold line). Figure (a) shows simulated data with white noise (SNR=20); Figure (b) shows the estimated curve using the new method; Figure (c) demonstrates the standard single kernel regression result.

3.2. Application to real data: Burning Ethanol Data

To evaluate the performance of our proposed method in practice, we analyzed the Burning Ethanol Data set. Figure 2(a) shows the data set of Brinkmann (1981) that has been analyzed extensively. The data consist of 88 a measurement from an experiment in which ethanol was burned in a single cylinder automobile test engine. Because of the nature of the experiment, the observations are not available at equally-spaced design points, and the variability is larger for low equivalence ratio.

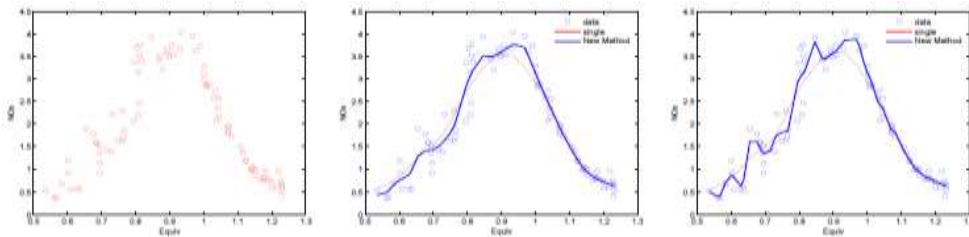


Fig. 2. Figure (a) shows Burning Ethanol Data; The blue bold lines in figure (b) and (c) show different estimated curves with two and three kernels.

As we all know, it is a difficult problem to control the pump around 0.8. Figure 2 shows the regression results with different parameters and iteration steps. The red line represents the single kernel estimator. The blue in figure 4(b)-(c) represent the estimators after two and three steps iteration with different kernel bandwidths which are determined by GCV. From the experimental results, several advantages can be drawn. First, all the estimated curves have not a spurious high-frequency feature when the equivalence ratio is around 0.8 which is the drawback other regression methods must deal with cautiously. Second, compared with [13], the proposed method is not sensitive to the pilot estimator and the kernel bandwidth selection. Finally, all the fitting results show the good boundary performance.

4. Conclusions and discussion

In this paper, we consider the kernel trick and proposed an iterative localized reweighted kernel regression method which includes two parts. At first, an improved Nadaraya-Watson estimator is introduced based on blockwise approach, which improves the classical Nadaraya-Watson estimator to be adaptive to different portions wit different curvatures; Then, considering the shortcoming of general MKR, we proposed a novel kernel selection framework during iteration procedure which avoids underfitting and overfitting effectively.

The simulation results show that the proposed method is less sensitive to noise level and pilot selection. Furthermore, experiments on simulated and real data set demonstrate that the new method is adaptive to the local curvature variation and improves boundary performance. It is easy to extend the method to other type additive noise. Kernel function plays an important role in kernel trick and can only work well in some circumstances, so, how to construct a new kernel function according to the given sample data settings is another direction we will keep up with.

References

- [1] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol.20, pp. 2626-2635, 2004.
- [2] D. Zheng, J. Wang and Y. Zhao, "Non-flat function estimation with a multi-scale support vector regression," *Neurocomputing*, vol. 70, pp. 420-429, 2006.
- [3] B. Scholkopf and A. J. Smola, *Learning with Kernels*. London, England: The MIT Press, Cambridge, Massachusetts, 2002.
- [4] M. Gonen and E. Alpaydin, "Localized multiple kernel learning," in *Processing of 25th International Conference on Machine Learning*, 2008.
- [5] M. Szafranski, Y. Grandvalet and A. Rakotomamonjy, "Composite kernel learning," in *Processing of the 25th International Conference on Machine Learning*, 2008.
- [6] G. R. G. Lanckriet, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27-72, 2004.
- [7] A. Rakotomamonjy, F. Bach, S. Canu and Y. Grandvalet, "More efficiency in multiple kernel learning," *Proceedings of the 24th international conference on Machine Learning*, vol. 227, pp. 775-782, 2007.
- [8] E. A. Nadaraya, "On estimating regression," *Theory of probability and Its Applications*, vol. 9, no. 1, pp. 141-142, 1964.
- [9] G. S. Watson, "Smooth regression analysis," *Sankhya, Ser. A*, vol. 26, pp. 359-372, 1964.
- [10] Y. Kim, J. Kim and Y. Kim, "Blockwise sparse regression," *Statistica Sinica*, vol. 16, pp. 375-390, 2006.
- [11] L. Lin, Y. Fan and L. Tan, "Blockwise bootstrap wavelet in nonparametric regression model with weakly dependent processes," *Metrika*, vol. 67, pp. 31-48, 2008.
- [12] A. Tikhonov and V. Arsenin, *Solutions of Ill-posed Problem*, Washington: W. H. Winston, 1977.
- [13] A. Rakotomamonjy, X. Mary and S. Canu, "Non-parametric regression with wavelet kernels," *Applied Stochastic Models in Business and Industry*, vol. 21, pp. 153-163, 2005.