Adaptive Fit Parameters Tuning with Data Density Changes in Locally Weighted Learning

Han Lei, Xie Kun Qing, and Song Guo Jie*

Key Laboratory of Machine Perception (Ministry of Education), Peking University
{hanlei,kunqing}@cis.pku.edu.cn, gjsong@pku.edu.cn

Abstract. Locally weighted learning (LWL) is a form of lazy learning and focuses on locally weighted regression. Due to its high efficiency and flexibility, the learning mechanism is widely used in prediction. However, LWL fails when the data points are sparse, and fewer survey concerns about tuning fit parameters in local model with density of the data input. This paper discusses the relationship between data density and fit parameters from a theoretical view. The relationship we advocate also contributes to adaptive fit parameters selection. Experimental studies provide evidence for the mathematical derivation and show its application superiority in prediction of traffic flow.

Keywords: Density; Locally weighted learning; Adaptive; Tuning.

1 Introduction

Locally weighted learning contains several important parts known as local model structures, weighting functions, smoothing parameters, distance functions etc. Important fit parameters such as bandwidth h and number of neighbors K are discussed by Atkeson et al. (1997) [1]. When robustness is emphasized, this learning mechanism is required to allow adaption to changes in data density and distribution.

However, locally weighted learning fails or performs badly when the data density is very low. Low data density makes the local regression powerless because of the barren neighborhood around the query. This difficulty can be overcome by tuning the fit parameters adaptively in prediction.

Take a smoothing or bandwidth parameter h as an example, many adaptive selection methods are proposed such as Partially Adaptive Bandwidth [2]. Global Bandwidth Selection(QBS), providing a global optimal bandwidth, is widely used due to its simplicity and universality, but collapses with changes in data density and noise. Query-based Bandwidth Selection(QBS) and Point-based Bandwidth Selection(PBS) using a bandwidth associated with each data or query point will allow rapid or asymmetric changes in the behavior of the data, but they use a lot of storage associated with data points and require expensive calculations in preprocessing as well as updating.

[*] Corresponding Author.

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part II, LNCS 6064, pp. 408–415, 2010.

This paper discusses the relationship between data density and fit parameters in local constant model from a theoretical view. Strict mathematical derivation of this relationship is evolved to explain the process of relation detection. The results produce a theorem which reflects the relationship of data density and two fit parameters in the local model – bandwidth h and number of neighbors K. Experiments are divided into two groups. The first experiment based on artificial stochastic dataset is implemented to verify the correctness of the theorem. Then the theory is used in the prediction of traffic flow to show its superiority.

In this paper, the first section is an introduction. The second part describes the local constant model and its components. Section three is the mathematical derivation of relation detection in detail. Experimental studies are described in the forth section. Section five makes a conclusion.

2 Preliminaries and Problem Statement

The learning problems have a standard regression model:

$$\mathbf{y} = f(x) + \varepsilon \tag{1}$$

where $\boldsymbol{x}$ denotes the N-dimensional input vector, $\boldsymbol{y}$ the scalar output, and ε a mean-zero noise item. When the regression model is used for prediction of time series, $\boldsymbol{x}$ is a vector with lagged values of $\boldsymbol{y}$. The local regression is used to approximate the unknown function $\boldsymbol{f}()$. A constant local model is represented by the following prediction equation:

$$\hat{y}(q) = \frac{\sum_i^K y_i G(\frac{d(x_i,\mathbf{q})}{h})}{\sum_i^K G(\frac{d(x_i,\mathbf{q})}{h})} \tag{2}$$

$\mathbf{q}$ is the query. $d()$ is a distance function usually in a typical formation as the Euclidean distance, $d(\mathbf{x},\mathbf{q}) = \sqrt{\sum_i^K (x_i - q_i)^2}$. A weighting or kernel function G() is used to calculate a weight for data points from the distance. A typical weighting function is Gaussian, $G(d) = e^{-d^2}$. A bandwidth h defines the scale or range over which generalization is performed. K denotes the number of the nearest neighbors that take part in local regression.

Based on the theory above, data density should be introduced to the regression with local constant model together. When a given query is located in sparseness, the barren neighborhood makes the constant model structure in equation 2 powerless because of small weights of neighbors if bandwidth h does not make any corresponding adjustments. Note that there is no variable which refers to density in local constant model, it is difficult to find any adaptive correspondences between fit parameters and data density. Therefore, the following derivation process is aimed to overcome this problem.

3 Fit Parameters Tuning with Density Changing

We first formalize the data density for the following theory derivation.

3.1 Expression of Local Constant Model with Data Density

Data density, reflecting the denseness or sparseness of data distribution, is represented as:

$$\rho = \frac{K}{V} \quad in \quad 3D \quad \textbf{or} \quad \rho = \frac{K}{S} \quad in \quad 2D \tag{3}$$

K is the number of data points in a local circular region which is measured by volume V or area S. Take a planar region as example, $S = \pi r^2$, where r_k is a radius of the circular region. If we treat this circle as the neighborhood of a query point $\mathbf{q}$ with K neighbors and assume that $\mathbf{q}$ is the centre of the circular region, the circle with radius r_k can be regarded as a local model.

In equation 2, r_k is also the distance between the Kth neighbor and the query. That means the Kth neighbor is located on the circumference of the circle. Under a critical localized limitation (which is a constraint for selection of the number of neighbors K), the local region is assumed to be symmetrically distributed and the density ρ a constant. Thus, each neighbor i has its own distance r_i to the centre query, and this distance can be represented as:

$$d(x_i, \mathbf{q}) = r_i = \sqrt{\frac{i}{\rho \times \pi}} \tag{4}$$

Inserting equation 4 in equation 2, we obtain:

$$\hat{y}(q) = \frac{\sum_i^K y_i G(\sqrt{\frac{i}{\rho \pi h^2}})}{\sum_i^K G(\sqrt{\frac{i}{\rho \pi h^2}})} \tag{5}$$

3.2 Error Function

The error of estimation is denoted as:

$$e = |\hat{y}(\mathbf{q}) - y| = \left| \frac{\sum_i^K y_i G(\sqrt{\frac{i}{\rho \pi h^2}})}{\sum_i^K G(\sqrt{\frac{i}{\rho \pi h^2}})} - y \right| = \left| \frac{\sum_i^K (y_i - y) G(\sqrt{\frac{i}{\rho \pi h^2}})}{\sum_i^K G(\sqrt{\frac{i}{\rho \pi h^2}})} \right| \tag{6}$$

This error equation implicates the relationship of number of neighbors K, bandwidth h and data density ρ. However, the term $y_i - y$ is a stochastic entry according to the actual input values. In order to facilitate analysis, the stochastic entry should be changed into non-stochastic variables. Actually, item $y_i - y$ is determined by the input which has an explicit and fixed distribution. Thus, in our assumed symmetrical local region, $y_i - y$ is constrained by ρ and stochastic oscillates in a limited range. We introduce a standard data distribution S_0 with a constant density ρ_0 and a query $\mathbf{q}_0$ in it. For any dataset S_t with ρ_t and $\mathbf{q}_t$, there is a relationship reflected in the distance r_{0k} and r_{tk}:

$$\frac{r_{tk}}{r_{0k}} = \frac{\sqrt{\frac{K}{\rho_t \times \pi}}}{\sqrt{\frac{K}{\rho_0 \times \pi}}} = \sqrt{\frac{\rho_0}{\rho_t}} \tag{7}$$

As $\boldsymbol{y}$ is the output of input $\boldsymbol{x}$, when $\hat{y}$ is used as a prediction for time series, $\boldsymbol{y}$ is actually the next state of $\boldsymbol{x}$. Due to Local Dependence Rule [3], $\boldsymbol{y}$ holds the similar features as $\boldsymbol{x}$. Thus, output $\boldsymbol{y}$ has the similar distribution as $\boldsymbol{x}$:

$$\frac{y_{ti} - y_t}{y_{0i} - y_0} \approx \frac{x_{ti} - q}{x_{0i} - q} \tag{8}$$

where $x_i - \mathbf{q}$ denotes one of the feature distances in multi-dimensional input cases, and it is positively related to data point distance r_i. We represent this positive correlation as:

$$x_i - q = T(r_i) \tag{9}$$

T() has various formats that depends on the features of the input data points ($|x_i - q| = r_i$ when $\boldsymbol{x}$ is a single-dimensional time serie). We assume T() as a linear formation for simplification. Then we can get from equation 7, 8 and 9:

$$\frac{y_{ti} - y_t}{y_{0i} - y_0} \approx \frac{x_{ti} - q}{x_{0i} - q} = \frac{T(r_{ti})}{T(r_{0i})} = \frac{r_k^t}{r_k^0} = \sqrt{\frac{\rho_0}{\rho_t}} \tag{10}$$

and equation 6 is changed by inserting equation 10:

$$e = \left| \frac{\sum_i^K \sqrt{\frac{\rho_0}{\rho_t}} \cdot (y_{0i} - y_0) \cdot G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)}{\sum_i^K G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)} \right| \tag{11}$$

To completely eliminate the stochastic entry, we assume again that T() has a squared formation (If T() keeps identical with equation 10 as a linear formation, we get Integral Error Function, see in Section 3.3).

$$e = \left| \frac{\sum_i^K \sqrt{\frac{\rho_0}{\rho_t}} \cdot C \cdot {r_{0i}}^2 \cdot G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)}{\sum_i^K G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)} \right| = \left| \frac{\sum_i^K \frac{C \cdot i}{\pi \sqrt{\rho_t \rho_0}} \cdot G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)}{\sum_i^K G\left(\sqrt{\frac{i}{\rho_t \pi h^2}}\right)} \right| \tag{12}$$

Treat error e as Error Function E() about ρ, K and h using a Gaussian Kernel, equation 12 can be rewrote as:

$$E(\rho, K, h) = \left| \frac{\sum_i^K \frac{C \cdot i}{\pi \sqrt{\rho \rho_0}} \cdot e^{-\frac{i}{\rho \pi h^2}}}{\sum_i^K e^{-\frac{i}{\rho \pi h^2}}} \right| \tag{13}$$

C is a temp coefficient and will be found useless in the following part.

3.3 Boundary Determination of Error Function

Theorem: By constraining the boundary of error function E(), we propose a theorem called Relation of Density and Parameters Theorem (RDPT):

$$\rho = \frac{K}{\pi h^2}$$

Proof. Integral scaling techniques based on Squeezing Theorem are used here for derivation. Exponential function $y = e^{-x/a}$ has some properties that the sum of exponential terms with integer x from 1 to K is larger than its integral from 1 to K, and the sum of exponential terms with integer x from 2 to K is fewer than its integral from 1 to K.

$$\sum_{i=1}^{K} e^{-\frac{i}{\rho\pi h^2}} > \int_1^K e^{-\frac{i}{\rho\pi h^2}} di = -\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}$$

$$\sum_{i=1}^{K} e^{-\frac{i}{\rho\pi h^2}} < e^{-\frac{1}{\rho\pi h^2}} + \int_1^K e^{-\frac{i}{\rho\pi h^2}} di = -\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} + e^{-\frac{1}{\rho\pi h^2}}$$

Propose a variable B_1, and set it to:

$$\rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} < B_1 < \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}} + e^{-\frac{1}{\rho\pi h^2}}$$

The similar properties happened in function $y = xe^{-x/a}$.

$$\sum_{i=1}^{K} \frac{C \cdot i \cdot e^{-\frac{i}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} > \frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \left(\left(-K - \rho\pi h^2\right) \cdot e^{-\frac{K}{\rho\pi h^2}} + \left(1 + \rho\pi h^2\right) \cdot e^{-\frac{1}{\rho\pi h^2}} \right)$$

$$\sum_{i=1}^{K} \frac{C \cdot i \cdot e^{-\frac{i}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} < \frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \left(\left(-K - \rho\pi h^2\right) \cdot e^{-\frac{K}{\rho\pi h^2}} + \left(1 + \rho\pi h^2\right) \cdot e^{-\frac{1}{\rho\pi h^2}} \right) + \frac{C \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}}$$

B_2 is proposed here like B_1 and set to:

$$\frac{C \cdot \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} \cdot \left(1 + \rho\pi h^2\right) < B_2 < \frac{C \cdot \rho\pi h^2 \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}} \cdot \left(1 + \rho\pi h^2\right) + \frac{C \cdot e^{-\frac{1}{\rho\pi h^2}}}{\pi\sqrt{\rho\rho_0}}$$

Introduce B_1 and B_2 into error function:

$$E(\rho, K, h) = \frac{\frac{C \cdot \rho\pi h^2}{\pi\sqrt{\rho\rho_0}} \cdot e^{-\frac{K}{\rho\pi h^2}} \cdot \left(-K - \rho\pi h^2\right) + B_2}{-\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} + B_1} \tag{14}$$

Note that if T() in Section 3.2 keeps identical with equation 10 as linear, we get Integral Error Function which is twice the integral of the Gaussian distribution with 0 mean and variance of 1/2, $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, and this Integral Error Function cannot be derived to a non-integral formula. In order to facilitate analysis, we select two difference formations of T(), and both of the two can reflect the positive correlation of $x_i - q$ and r_i which is the most important.

Because most of the errors are generated in prediction of points in sparseness, the core focus should be put on low density situation. When data density is under a threshold ρ_{low}, both B_1 and B_2 approach zero. Then we can do an approximation:

$$E(\rho, K, h) \approx \frac{\frac{C}{\pi\sqrt{\rho\rho_0}} \cdot \rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}} \cdot \left(-K - \rho\pi h^2\right)}{-\rho\pi h^2 \cdot e^{-\frac{K}{\rho\pi h^2}}} = \frac{CK}{\pi\sqrt{\rho\rho_0}} + \frac{C\sqrt{\rho}h^2}{\sqrt{\rho_0}}, \quad \rho \leq \rho_{low}$$

To minimize this error function, we calculate its derivative about ρ and make the derivative equal to zero:

$$\frac{\partial E(\rho, K, h)}{\partial \rho} = -\frac{1}{2} \cdot \frac{CK}{\pi\sqrt{\rho_0}} \cdot \rho^{-\frac{3}{2}} + \frac{1}{2} \cdot \frac{Ch^2}{\sqrt{\rho_0}} \cdot \rho^{-\frac{1}{2}} = 0 \tag{15}$$

Finally, we get the compendious relationship of ρ, K and h, named Relation of Density and Parameters Theorem:

$$\rho = \frac{K}{\pi h^2} \tag{16}$$

4 Experiments

Experiments are designed into two ways: (1) Verification by artificial stochastic datasets; (2) Application of RDPT theory in prediction of traffic flow. Experiments are implemented in Matlab 7, Intel Core II 2.26G CPU and 2G memory.

4.1 Verification by Artificial Stochastic Datasets

To prove the theoretical results above by data points, experimental studies are designed to test the theorem. The studies described in this part are based on stochastic datasets with variational local density. Those datasets are generated by random numbers whose elements are normally distributed with mean 0, variance $\sigma^2 = 1$ and standard deviation $\sigma = 1$, shown in figure 1. The size of sample input $\boldsymbol{x}$, query input $\mathbf{q}$, sample output $\boldsymbol{y}$ and query output $\boldsymbol{y}_q$ are M. $\boldsymbol{x}$ and $\mathbf{q}$ is generated by the stochastic process while $\boldsymbol{y}$ and $\boldsymbol{y}_q$ is generated by adding a random noise term to $\boldsymbol{x}$ and $\mathbf{q}$ separately in order to simulate real time series whose outputs hold the similar features with inputs. The prediction accuracy is measured by MFE:

$$MFE = \frac{1}{M} \sum_{i=1}^{M} \frac{|y_{qi} - \hat{y}_{qi}|}{y_{qi}}$$

Figure 2 shows the comparison of two relations when ρ is lower than a threshold ρ_{low}(Set to be the maximum density of the 20% sparsest data points). The blue curve shows the relation of Optimal Bandwidths, calculated by ideally minimizing every query prediction errors with a "prescient" knowledge of real output(OBS), and data density calculated by Kernel Density Estimation [4]. The red one shows the $\rho = K/\pi h^2$ relation curve. K is constrained to be 10 here. The results show a good match of RDPT to optimal bandwidths.

Table 1 shows the prediction MFE of several bandwidth selection methods. Global Bandwidth Selection doesn't need any preprocessing and additional storage besides samples, but its performance is the worst. Query-based and Point-based Bandwidth Selection are both calculating expensive in preprocessing, and need additional storage depended on size of query and sample separately, but perform better than Global Bandwidth Selection. RDPT we proposed needs no additional storage and show the best results.

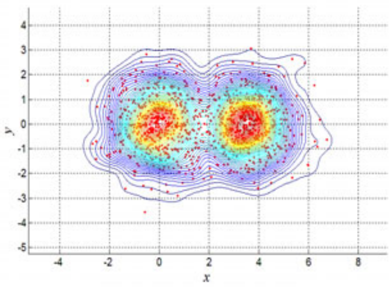

Fig. 1. Stochastic Input with Density Contour Lines

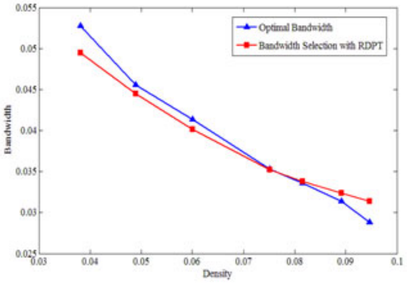

Fig. 2. Comparison of Optimal Relation and RDPT Relation, $\rho <= \rho_{low}$

Table 1. Comparison of Prediction Mean Fractional Error

M	Bandwidth Selection	MFE	Time consumed in prediction(s)	Time consumed in preprocessing(s)	Additional storage consumed in prediction
1000	GBS	0.1298	0.7330	0	0
1000	QBS	0.1058	0.7950	16.2710	1000
1000	PBS	0.1292	0.7960	16.0360	1000
1000	RDPT	0.0910	0.7170	0.1400	0
1000	OBS	0.0669	-	-	-

The RDPT policy is a combination with global bandwidth selection. When $\rho <= \rho_{low}$, the RDPT method is used, and global bandwidth selection is implemented when $\rho > \rho_{low}$. Although, data points in sparseness are fewer than points in dense fields, but large errors are taken place in sparseness which determines mostly the quality of entire prediction. Therefore, a combination mechanism is a good choice.

4.2 Application in Prediction of Traffic Flow

Experiment implemented in this section is an application of RDPT policy in prediction of traffic flow. The studies are based on Freeway Performance Measurement System (PeMS) from Berkeley University of California [5]. Traffic datasets are from Los Angeles mainline I5, segment 759700, 759707 and 716936. The data points are collected every 5 minutes interval including states of traffic flow and occupancy. Sample data points are selected from Oct.1 2009 to Oct.16 2009 exclude weekends and holidays. Query points are from Oct.19 2009.

Both Global Bandwidth Selection(GBS) and RDPT Bandwidth Selection are used in prediction of traffic flow. Figure 3 and figure 4 show the two comparisons of prediction and real output. From table 2 and comparison of figure 3 to figure 4, it can be found that our RDPT adaptive bandwidth selection shows its great advantages in traffic flow. Both of the two experiments afford sufficient evidence for RDPT theory and show its comprehensive application.

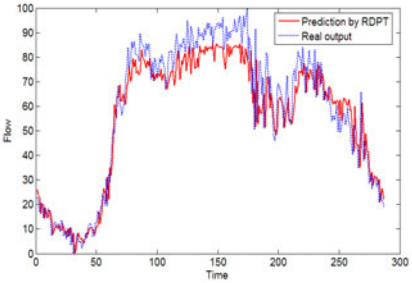

Fig. 3. Prediction with GBS

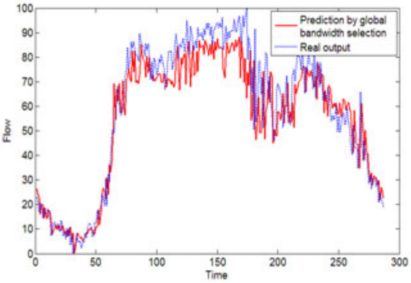

Fig. 4. Prediction with RDPT

Table 2. Comparison of Prediction Mean Fractional Error

Segment No.	MFE of GBS	MFE of RDPT
759700	0.1254	0.1134
716907	0.1059	0.0996
716936	0.1346	0.1057

5 Conclusions

The relationship of data density and fit parameters is discussed in this paper. After preliminaries, detail mathematical derivation of the RDPT theory is proposed. Experimental studies are implemented to prove its correctness and to show its advantages in prediction of traffic flow.

In this paper, there is a main assumptions: two formations of T() described in 3.2 and 3.3. This assumptions can be extended or modified in future research about linear regression model and other model structures. High-dimensional regression [6] should also be discussed in future work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 60703066 and No. 60874082; supported by the National High-Tech Research and Development Plan of China (863) under Grant No.2006AA12Z217.

References

1. Atkeson, C., Moore, A., Schaal, S.: Locally Weighted Learning. Artificial Intelligence Review, 11–73 (1997)
2. Grabis, J.: Partially Adaptive Bandwidth Used in Prediction with Local Regression. Riga Technical University, Kalku 1, Riga Lv-1658, Latvia
3. Christopher, D.L.: Local Models for Spatial Analysis. Queen's University (2007)
4. Kernel Density Estimation, `http://en.wikipedia.org/wiki/`
5. Freeway Performance Measurement System, `http://pems.eecs.berkeley.edu`
6. Vijayakumar, S., Souza, A.D., Schaal, S.: Incremental Online Learning in High Dimensions. Neural Computation 17 (2005)