Journal of the Korean Data & Information Science Society
2012, **23**(3), 579–585

http://dx.doi.org/10.7465/jkdi.2012.23.3.579
한국데이터정보과학회지

Study on semi-supervised local constant regression estimation[†]

Kyungha Seok[1]

[1]Department of Data Science, Inje University

Received 7 March 2012, revised 3 April 2012, accepted 10 April 2012

Abstract

Many different semi-supervised learning algorithms have been proposed for use with unlabeled data. However, most of them focus on classification problems. In this paper, we propose a semi-supervised regression algorithm called the semi-supervised local constant estimator (SSLCE), based on the local constant estimator (LCE), and reveal the asymptotic properties of SSLCE. We also show that the SSLCE has a faster convergence rate than that of the LCE when a well chosen weighting factor is employed. Our experiment with synthetic data shows that the SSLCE can improve performance with unlabeled data, and we recommend its use with the proper size of unlabeled data.

Keywords: Convergence rate, local polynomial regression, Nadaraya Watson estimator, semi-supervised regression, smoothing parameter, weighting factor.

1. Introduction

In many practical applications such as speech recognition, email classification and text classification, there is often a large amount of unlabeled data available. However, labeling this data is expensive, difficult, and time consuming. The value of the unlabeled data was not clearly understood, bringing about a rise in the study of semi-supervised learning (SSL) starting in the mid-1990s. As the demand for unlabeled data has increased, SSL has become increasingly important as an analysis tool.

The promising empirical success of SSL algorithms in favorable situations has triggered several recent attempts (Lafferty and Wasserman, 2008; Niyogi, 2008) at developing a theoretical understanding of SSL. In a recent paper, Singh *et al.* (2008) established that if the complexity of the distributions under consideration is too high to be understood using labeled data points, but is small enough to be understood using unlabeled data points, using a finite sample analysis in SSL can improve the performance of a supervised learning task. There have been many successful practical SSL algorithms generated as summarized in Chapelle *et al.* (2006), Sindhwani (2005), Zhu (2005), Xu *et al.* (2010) and Zhu and Goldberg (2009).

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0009705).

[1] Professor, Department of Data Science and Institute of Statistical Information, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: statskh@paran.com

It is worthwhile noting that the previous research focused primarily on classification. According to the work of Belkin *et al.* (2006) and Zhu (2005), a graph-based method could be applied to the regression estimator. This means that unlabeled data can contain helpful information and can increase the performance of the regression estimator. Zhou and Li (2005) proposed a semi-supervised regression (SSR) algorithm named COREG that boosts regression accuracy by exploiting unlabeled data in two k-nearest neighbor regression estimators using different distance metrics, each of which labels the unlabeled data for the other regression estimator. Wang *et al.* (2006) developed an SSR algorithm called semi-supervised kernel regression (SSKR) based on the classical kernel regression estimator. A weighting factor is used to fine-tune the effect of the unlabeled data in the SSKR. They also investigate the connection between the SSKR and the graph-based method. They showed that with a properly-chosen weighting factor, the SSKR remarkably outperformed kernel regression and the graph-based method. Cortes and Mohri (2007) dealt with regression problems in a transductive setting. They gave a new error boundary for transductive regression that holds for all bounded loss functions and coincides with the tight classification bounds of Vapnik (1998). Based on the given error bound, they presented a new algorithm for transductive regression that performs well and can scale to large data sets. Xu *et al.* (2011) proposed a semi-supervised least squares support vector regression and showed their feasibility and efficiency by experiment on a corn data set.

Some existing SSR methods have empirically shown promising performance. However, Lafferty and Wasserman (2008) showed that SSR methods based on regularization using graph Laplacians do not lead to faster minimax rates of convergence than those of kernel regression estimators. They also revealed that the estimators using unlabeled data do not have smaller risks than the estimators that use only labeled data when the dimension of the input variable is greater than 1. In this paper, we derive an SSLCE from the LCE view point for the univariate input variable. The derived estimator is the same as the SSKR estimator proposed by Wang *et al.* (2006). However, we reveal the asymptotic properties of the SSLCE. From the asymptotic properties, we know that with a properly-chosen weighting factor, we can obtain an SSLCE that has a faster convergence rate than LCE $O(n^{-4/5})$. From the numerical experiment and theoretical analysis, we also recommend using a proper quantity of unlabeled data as a larger amount of unlabeled data does not guarantee better performance for the SSLCE. We have confirmed this proposed assertion through numerical study.

The rest of this paper is organized as follows. Section 2 reviews the LCE. Section 3 introduces the SSLCE. Section 4 presents some asymptotic properties of SSLCE and section 5 gives some numerical results. Section 5 gives our conclusion and discussion for further study.

2. Local constant estimator

It was shown extensively in the literature that the local polynomial approximation method has various nice features (Fan and Gijbels, 1996). Assume that the bivariate data $(X_1, Y_1), \cdots,$ (X_{n_L}, Y_{n_L}) is an i.i.d sample from the model

$$Y = m(X) + \sigma(X)\epsilon,\ E(\epsilon) = 0, Var(\epsilon) = 1$$

where X and ϵ are independent random variables. The objective is to estimate the regression function $m(x) = E(Y|X = x)$ based on the observations $(X_1, Y_1), ..., (X_{n_L}, Y_{n_L})$. The marginal density of X will be denoted by $f_X(\cdot)$. The above location-scale model is not a necessity, but a convenient way to introduce the notations. Suppose that the $(p+1)^{th}$ derivative of $m(x)$ at the point x_0 exists. We then approximate $m(x)$ locally by a polynomial of order p :

$$m(x) \sim m(x_0) + m'(x_0)(x - x_0) + \cdots + m^{(p)}(x_0)(x - x_0)^p/p! \tag{2.1}$$

for x in a neighborhood of x_0, and do a local polynomial regression fit

$$min_\beta \sum_{i=1}^{n_L} \{Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j\}^2 K_{h_L}(X_i - x_0) \tag{2.2}$$

where $\beta = (\beta_0, \cdots, \beta_p)'$. Here $K_h(\cdot)$ denotes a nonnegative kernel function with smoothing parameter h_L which determines the size of the neighborhood of x_0. Let $\{\hat{\beta}_\nu\}$ denote the solution to the weighted least squares problem (2.2). Then it is obvious from the Taylor expansion (2.1) that $\nu!\ \hat{\beta}_\nu$ estimates $m^{(\nu)}(x_0)$, $\nu = 0, \cdots, p$. When $p = 0$, the objective function (2.2) becomes

$$min_{\beta_0} \sum_{i=1}^{n_L} \{Y_i - \beta_0\}^2 K_{h_L}(X_i - x_0) \tag{2.3}$$

and the solution $\hat{\beta}_0 = \hat{m}_L(x_0)$ is

$$\hat{m}_L(x_0) = \frac{\sum_{i=1}^{n_L} Y_i K_{h_L}(X_i - x_0)}{\sum_{i=1}^{n_L} K_{h_L}(X_i - x_0)}. \tag{2.4}$$

The above LCE in (2.4) is the same as the Nadaraya-Watson estimator which is independently proposed by Nadaraya (1964) and Watson (1964) and one of the most popular methods in kernel regression.

3. Semi-supsevised local constant estimator

In this section we introduce the formulation of SSLCE. To utilize the unlabeled data set $U = \{X_{n_L+1}, \cdots, X_n\}$, where $n = n_U + n_L$ is total size of the labeled and unlabeled data, we modified the objective function as

$$min_{\beta_0} \sum_{i=1}^{n_L} (Y_i - \beta_0)^2 K_{h_L}(X_i - x_0) + \lambda \sum_{j=n_L+1}^{n} (\hat{Y}_j - \beta_0)^2 K_{h_U}(X_j - x_0), \tag{3.1}$$

where λ is a weighting factor to modulate the labeled and unlabeled data, and h_U is a smoothing parameter used in unlabeled data. $\hat{Y}_j$ is a pilot estimator of $m(X_j)$ for $j = n_L + 1, \cdots, n$. In this paper we used LCE as the pilot estimator. The solution of (3.1) is the SSLCE

$$\hat{m}_{SS}(x_0) = \frac{\sum_{i=1}^{n_L} Y_i K_{h_L}(X_i - x_0) + \lambda \sum_{j=n_L+1}^{n} \hat{Y}_j K_{h_U}(X_j - x_0)}{\sum_{i=1}^{n_L} K_{h_L}(X_i - x_0) + \lambda \sum_{i=n_L+1}^{n} K_{h_U}(X_j - x_0)} \tag{3.2}$$

The estimator (3.2) is the same as the one proposed by Wang *et al.* (2006) but they used the same value of h_L and h_U. According to the their work, the estimator is closely related with graph-based method and provide another viewpoint for graph-based method. They showed that it remarkably outperforms kernel regression and the graph based method.

4. Asymptotic properties

To proceed the asymptotic properties, we need the follwing conditions:

1. The regression function $m(x)$ has a bounded and continuous second derivative.
2. The conditional variance $\sigma^2(x) = var(Y|X = x)$ is bounded and continuous.
3. The marginal density f_X of the covariate X is continuous.
4. The kernel function K is bounded density function with $\int_{-\infty}^{\infty} xK(x)dx = 0$ and $\int_{-\infty}^{\infty} x^4 K(x)dx \leq \infty$

In the sequel we denote $\mu_i = \int_{-\infty}^{\infty} u^i K(u)du$, $\nu_i = \int_{-\infty}^{\infty} u^i K^2(u)du$. We state the following pointwise and global properties of the SSLCE and omit their proofs.

Theorem 4.1 Under the conditions 1-4, if $h_L \to 0$ and $n_L h_L \to \infty$, then LCE $\hat{m}_L(x)$ has the asymptotic distribution

$$\hat{m}_L(x) - m(x) \to N(h_L^2 \mu_2 B(x), \frac{\sigma^2(x)\nu_0}{n_L h_L f_X(x)}) \tag{4.1}$$

where $B(x) = \frac{1}{2}m''(x) + m'(x)f_X'(x)/f_X(x)$.

Theorem 4.2 Under the conditions 1-4, if $h_U \to 0$ and $n_U h_U \to \infty$, SSLCE $\hat{m}_{SS}(x)$ has the asymptotic distribution

$$\hat{m}_{SS}(x) - m(x) \to N(\frac{n_L h_L^3 + \lambda n_U h_U^3}{n_L h_L + \lambda n_U h_U}\mu_2 B(x), \frac{n_L h_L \sigma^2(x)\nu_0}{(n_L h_L + \lambda n_U h_U)^2 f_X(x)}). \tag{4.2}$$

Theorem 4.3 Under the conditions of Theorem 1, LCE $\hat{m}_L(x)$ has the asymptotic mean integrated squared error (AMISE)

$$AMISE(\hat{m}_L) = c_1 h_L^4 + \frac{c_2}{n_L h_L} \tag{4.3}$$

where $c_1 = \mu_2^2 \int B^2(x)dx$ and $c_2 = \nu_0 \int \frac{\sigma^2(x)}{f_X(x)}dx$.

Theorem 4.4 Under conditions of Theorem 2, SSLCE estimator $\hat{m}_{SS}(x)$ has the AMISE

$$AMISE(\hat{m}_{SS}) = c_1(\frac{n_L h_L^3 + \lambda n_U h_U^3}{n_L h_L + \lambda n_U h_U})^2 + \frac{c_2 n_L h_L}{(n_L h_L + \lambda n_U h_U)^2}. \tag{4.4}$$

From the Theorem 3 and 4 we know that $\hat{m}_{SS}$ can have faster convergence rate than that of $\hat{m}_L$ with a well chosen weighting factor. For examle if we choose λ such that $\mathrm{O}(n_L h_L^3) > O(\lambda n_U h_U^3)$ and $O(n_L h_L) < O(\lambda n_U h_U)$ then $AMISE(\hat{m}_{SS})$ is smaller than $AMISE(\hat{m}_L)$. This means that the unlabeled data can be used to make SSLCE have faster convergence rate. Since $n_L < n_U$ we may denote $n_U = n_L^p$, $p > 1$. In order to have faster convergence rate the above inequalities can be written as $O(\lambda n_L^{2p/5}) < O(n_L^{2/5})$ and $O(\lambda n_L^{4p/5}) < O(n_L^{4/5})$. So we can see $a < p < b$ for some a, b $(1 < a < b)$. This says that larger n_U does not guarantee better performance of SSLCE.

5. Numerical studies

Since Wang *et al.* (2006) has conducted several experiments for the SSLCE on many data sets, we can refer to their work on the behavior of the SSLCE. We have focused our study on the impact of the weighting factor and the size n_U of the unlabeled data.

The performance of the proposed SSLCE is illustrated through the use of simulated data. Note that the SSLCE in (3.2) is regarded as the weighted average of LCE with labeled data (LCEL) and with unlabeled data (LCEU) as follows:

$$\begin{aligned}\hat{m}_{SS}(x) &= \gamma \frac{\sum_{i=1}^{n_L} Y_i K_{h_L}(X_i - x)}{\sum_{i=1}^{n_L} K_{h_L}(X_i - x)} + (1-\gamma)\frac{\sum_{j=n_L+1}^{n} \hat{Y}_j K_{h_U}(X_j - x)}{\sum_{j=n_L+1}^{n} K_{h_U}(X_j - x)} \\ &= \gamma LCEL + (1-\gamma) LCEU,\end{aligned} \tag{5.1}$$

where $\gamma(x) (= \sum_{i=1}^{n_L} K_{h_L}(X_i - x)/(\sum_{i=1}^{n_L} K_{h_L}(X_i - x) + \lambda \sum_{j=n_L+1}^{n} K_{h_U}(X_j - x))$ is a weighting factor to modulate the labeled and unlabeled data. In this paper we let $\gamma(x) \equiv \gamma$. For simplicity, we can write $\mathbf{\hat{m}}_{ss} = (\hat{m}_{ss}(x_1), \cdots, \hat{m}_{ss}(x_n))' = Hy$, with $H = [\gamma K_L \ (1-\gamma)K_U]$, $y = (y_1, \cdots, y_n)'$ where $K_L = (K_{h_L}(X_i - X_j)/\sum_j K_{h_L}(X_i - X_j))_{n \times n_L}$, $i = 1, \cdots, n$, $j = 1, \cdots, n_L$ and $K_U = (K_{h_U}(X_i - X_j)/\sum_j K_{h_U}(X_i - X_j))_{n \times n_U}$, $i = 1, \cdots, n$, $j = n_L + 1, \cdots, n$. The parameters such as kernel parameters (h_L, h_U) and weighting parameter (γ) can be chosen through GCV function:

$$GCV(h_L, h_U, \gamma) = \frac{n \sum_{i=1}^{n} (Y_i - \hat{m}_{SS}(X_i))^2}{(n - trace(H))^2}.$$

Since the choice of parameters is beyond of this research, it is sufficient to compare LCEL with LCEU.

Example 5.1 To compare the performance of the LCEL with LCEU, we generated 100 data sets from the model $Y = sin(2\pi X) + e$, $X \sim U(0,1)$ and $e \sim N(0, \sigma^2)$. The sizes of labeled data $n_L = 50, 100, 200, 500$ and unlabeled data $n_U = 200, 500, 1000, 5000, 10000$ were used. The noise $\sigma = 0.1, 0.2$ was added to the model. To compare, we calculated the average of mean squared error (AMSE) and standard deviation of mean squared error (SDMSE) from the test data set of size 100.

We summarized the results in Figure 5.1. The average (connecting real line) and standard deviation (connecting dotted line) of mean squared error of LCEL (stars) and LCEU (circles) are presented in Figure 5.1 with $\sigma = 0.1, 0.2$, $n_L = 50, 100, 200, 500$ and $n_U = 200, 500, 1000, 5000, 10000$. The x-axis labels $1, \cdots, 5$ indicate $200, 500, 1000, 5000, 10000$ respectively. From Figure 5.1 we can determine that the LCEU has better performance than that of the LCEL for all cases except where $\sigma = 0.1$, $n_L = 500$ and $n_U > 500$. This means that we can always construct the SSLCE to behave better through the proper choice of γ. For $\sigma = 0.1$ and $n_L = 100$, any values of γ improve the SSLCE. In the case of $\sigma = 0.1$ and $n_L = 500$, the SSLCE is not worsened by choosing $\gamma = 1$.

Another important finding from Figure 5.1 was that a larger unlabeled data set is not more helpful to the SSLCE. Therefore, even though the size of the available unlabeled data is huge, we do not need to use all of it. When $\sigma = 0.2$ and $n_L = 50$, the size of the unlabeled data $n_U = 200$ yields better performance than that of when $n_U = 10000$ if the distributions of unlabeled data and labeled data are identical.

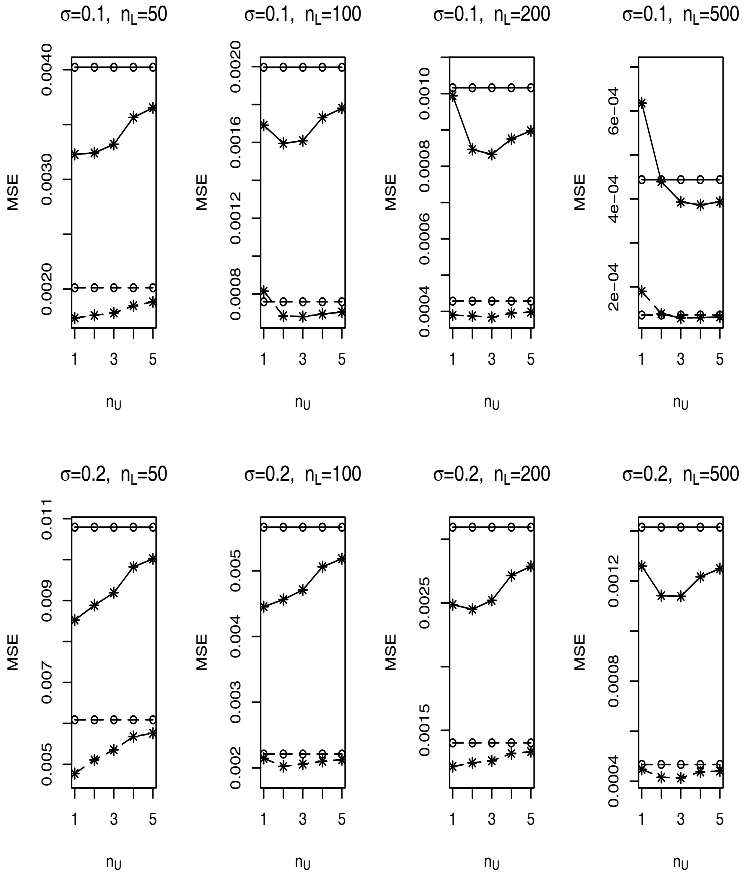

Figure 5.1 The average (connecting real line) and standard deviation (connecting dotted line) of mean squared error of LCEL (stars) and LCEU (circles)

6. Conclusions

In this paper, we derived an SSLCE from the LCE perspective for the univariate input variable. The derived estimator is the same as the SSKR estimator proposed by Wang *et al.* (2006). However we revealed the asymptotic properties of the SSLCE. From these asymptotic properties, we knew that with a properly-chosen weighting factor we could obtain an SSLCE that had a faster convergence rate than that of the LCE. From the numerical experiments we also recommended using the proper size unlabeled data set, as larger sizes of unlabeled data sets do not guarantee better performance for the SSLCE. We confirmed the proposed assertion through numerical study. The study of semi-supervised regression based on local polynomial regression and SSLCE with multivariate input variable will be considered in future work.

References

Belkin, M., Sindhwani, V. and Niyogi, P. (2006). Manifold regularization; A geometric framework for learning from examples. *Journal of Machine Learning Research*, **7**, 2329-2434.

Chapelle, O., Zien, A. and Scholkopf, B. (2006). *Semi-supervised learning*, MIT press, Boston.

Cortes, C. and Mohri, M. (2007). On transductive regression. *Advances in Neural Information Processing System*, **19**, 305-312.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, Chapman & Hall/CRC, London.

Hady, M. F. A. (2011). *Semi-supervised learning with committees*, Südwestdeutscher Verlag, für Hochschuischriften, Deutschland.

Lafferty, J. and Wasserman, L. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, **20**, 801-808.

Niyogi, P. (2008). *Manifold regularization and semi-supervised learning: Some theoretical analyses*, Technical Report TR-2008-01, CS Dept, U. of Chicago.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, 141-142.

Sindhwani, V., Niyogi, P. and Belkin, M. (2005). Beyond the point cloud: from transductive to semisupervised learning. In *ICML05, 22nd International Conference on Machine Learning*, 824 - 831.

Singh, A., Nowak, R. and Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems*, **21**, 1513-1520.

Vapnik, V. (1998). *The nature of statistical learning theory*, Springer-Verlag, New York.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, **26**, 359-372.

Wang, M., Hua, X., Song, Y., Dai, L. and Zhang, H. (2006). Semi-supervised kernel regression. In *Proceeding of the Sixth IEEE International Conference on Data Mining*, 1130-1135.

Xu, S., An. X., Qiao, X., Zhu, L. and Li, L. (2011) Semisupervised least squares support vector regression machines. *Journal of Information & Computational Science*, **8**, 885-892.

Xu, Z., King, I. and Lyu, M. R. (2010). *More than semi-supervised learning*, LAP LAMBERT Academic Publishing, London.

Zhou, Z. H. and Li, M. (2005). Semi-supervised regression with co-training. In *Proceeding of the 19th International Joint Conference in Artificial Intelligence*, 908-913.

Zhu, X. (2005). *Semi-supervised learning literature survey*, Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison.

Zhu, X. and Goldberg, A. (2009). *Introduction to semi-supervised learning*, Morgan & Claypool, London.