STUDIA INFORMATICA 2009
Volume 30 Number 3A (85)

Marcin MICHALAK
Politechnika Śląska, Instytut Informatyki

ADAPTIVE KERNEL ALGORITHMS FOR TIME SERIES PREDICTION

Summary. The article describes two kernel algorithms of the regression function estimation, that are used for the time series prediction. First of them (*HASKE*) has its own heuristic of the *h* parameter evaluation. The second (*HKSVR*) connects *SVM* and the *HASKE* in such way that it is based on the *HASKE* heuristic of local neighborhood evaluation.

Keywords: time series prediction, kernel estimators, nonparametric regression, support vectors machines

ADAPTACYJNE ALGORYTMY JĄDROWE W PREDYKCJI SZEREGÓW CZASOWYCH

Streszczenie. W artykule opisano dwa nowe algorytmy estymacji funkcji regresji, zastosowane do predykcji szeregów czasowych. Pierwszy z nich (*HASKE*) opiera się na pewnej heurystyce wyznaczania parametru wygładzającego. Drugi z nich (*HKSVR*) łączy *HASKE* z *SVR* przez wykorzystanie wspomnianej heurystyki.

Słowa kluczowe: predykcja szeregów czasowych, estymatory jądrowe, regresja nieparametryczna, maszyny wektorów podpierających

1. Introduction

Estimation of the regression function is the part of the machine learning discipline [28, 32]. In the general case estimation of the regression function consists of describing the unknown dependencies in the observed data set. Sometimes these relations can be intuitive or overt (obvious for the expert) or hidden.

Methods of the regression function estimation can be divided into two groups: parametrical, that consist of finding optimal values of the finite number of parameters, and nonparametrical, that don't assume any class of the function estimator.

The most of popular nonparametric regression estimators are spline functions [2, 10], radial basis functions [17], additive (and generalized additive) models [14], the *LOWESS* algorithm [6] or kernel estimators [22, 34] with support vector machines [3, 26].

Time series are the specific kind of data, because there is the time dependence between the consecutive samples. Methods of nonparametric regression function estimation can be also used for time series prediction, on the condition that the method of the smoothing parameter h had been modified [20].

This article describes two new methods of the time series prediction. Both of them are derived from kernel estimators. The first method is a kernel estimator (*HASKE*), which integral part is the heuristic of the h parameter evaluation. The second one is a hybrid algorithm that connects the mentioned kernel estimator and the model of support vector regression (*HKSVR*).

2. Regression function and its nonparametric estimators

Estimators, invented by the article author, belong to the group of nonparametrical regression function estimators. The following part of the article describes two important methods, that are the basis of the new regression models.

2.1. Kernel estimators

Kernel density estimator [18] is in general the function $\hat{f}:\mathbb{R}^m \to \mathbb{R}$ described in the following way:

$$\hat{f}(x) = \frac{1}{nh^m}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where n is a number of samples in the train set, K is the Borel function that satisfies the condition:

$$\int_{\mathbb{R}^m} K(x)dx = 1 \tag{2}$$

It is also assumed, that K function (called also as a kernel or kernel function) is symmetric with respect to zero and has weak global maximum at zero:

$$\begin{aligned} &\forall x \in \mathbb{R}^m \; K(x) = K(-x) \\ &\forall x \in \mathbb{R}^m \; K(0) \geq K(x) \end{aligned} \tag{3}$$

One of the most popular is the Epanechnikov kernel:

$$K(x) = \frac{3}{4}(1 - x^2) \quad for \; |x| \leq 1 \tag{4}$$

The *h* is called smoothing parameter and has very significant influence on estimator quality. On the basis of the kernel density estimators many of kernel regression estimators were developed. One of the most popular kernel estimator is Nadaraya-Watson estimator [22][34]:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} \tag{5}$$

Other popular kernel regression estimators are Gasser—Muller [13], Priestley—Chao [33], Stone—Fan [8].

As it is described in [25] the selection of the *h* parameter value is much more important than the choice of the kernel function. Small values of the *h* cause that the estimator fits data too much. Big values of this parameter *h* lead to the estimator that oversmooths dependencies in the analyzed set.

The most popular method of the *h* parameter evaluation is the analysis of the approximation of the Mean Integrated Square root Error. It leads to the following value of *h*:

$$h = 1,06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1,34}\right) n^{1/5} \tag{6}$$

where $\hat{\sigma}$ is the standard deviation from the train set and $\hat{R}$ means interquartile range from the train set. Details of derivations can be found in [25]. More advanced methods of *h* evaluation, for both: the density and regression estimation, can be found in [8, 12, 29, 30, 31].

2.2. Support vector machines

Support Vector Machines (*SVM*) were defined in [3] and later in [23, 32] as a tool for classification, but after some modifications it also can be used as an estimator of regression function. We assume, that the estimated function *f*(*x*) minimizes the following objective function:

$$J(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{7}$$

with constraints:

$$\begin{cases} y_i - wx_i - b & \leq & \varepsilon + \xi_i \\ wx_i + b - y_i & \leq & \varepsilon + \xi_i^* \\ \xi_i \,, \xi_i^* & \geq & 0 \end{cases} \tag{8}$$

The pair of the Lagrange multipliers $\alpha_i \,, \alpha_i^*$ is obtained for each train object. Then, the regression function is evaluated as:

$$\hat{f}(x, w) = w'x + w_0 \,, \tag{9}$$

where $w = \sum_{i=1}^{n} (\alpha - \alpha_i^*) x_i'$, and $w_0 = w'(x_r + x_s)/2$, where x_r, x_s are support vectors. Vector is called support when one of its Lagrange multipliers is not equal to zero. Detailed calculations can be found in [27]. Support Vector Regression (*SVR*) has also a number of its modifications, that use a local learning paradigm [9, 15].

3. Adaptive kernel estimators

Both of presented kernel estimators can be used as time series predictors, after the modification of the data space [20]. If the input space consists of pairs of samples (t, $x(t)$) then the output space consists of pairs (x_i, x_{i+pmax}) where p_{max} is the maximal prediction horizon. Fig. 1 shows the same time series (*G* from [4]) in two spaces.

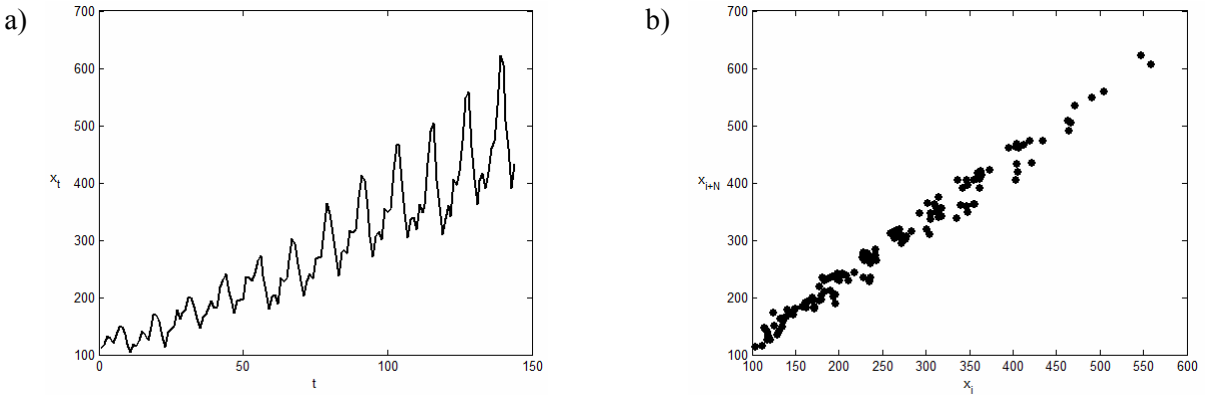

Fig. 1. The same time series (*G* series) shown in two different domains: a) the (t, $x(t)$) domain, b) the (x_i, x_{i+N}) domain ($N = 12$)

Rys. 1. Ten sam szereg czasowy (szereg *G*) przedstawiony w dwóch różnych dziedzinach: a) w dziedzinie (t, $x(t)$), b) w dziedzinie (x_i, x_{i+N}) ($N = 12$)

3.1. *HASKE* – adaptive kernel estimator

Let us define the set called *h*-neighborhood. If X is an explanatory variable the set of h-neighboors of point x_i is defined as:

$$h_{x_i} = \{x \in X : K(x, x_i) > 0\} \tag{10}$$

It is common, that for some extreme test points their *h*-neighborhoods are empty sets, what implies that the estimated value is 0 and the estimation error increases significantly. The solution of this problem is the *HASKE* algorithm (*Heuristic Adaptive Smoothing parameter Kernel Estimator*) [20]. The integral part of this estimator is the adaptive heuristic of evaluation the *h* parameter value. The author defines tune set as the subset of train set. If p_{max} means *p* last pairs from the train set, than these pairs mean the tune set. Author defines also the parameter μ, that modifies *h* in such a way: $h' = \mu h$. Starting from $\mu = 1$ this parameter is increased by the defined step as long as the regression error on the tune set decreases. Then as the optimal value of *h*' is concerned the value μh that gives the smallest regression error on the tune set. To assure that the μ value does not depend on the phase of the time series period, values μ_i for every phase are evaluated and μ becomes a median of all μ_i values ($err(t, p, h)$ is the estimation error of the *p* consecutive time series values, from the interval $[t+1, t+p]$, with the usage of *h* value of the smoothing parameter).

$$\mu = \text{med}\left\{\arg\min_m err(t-i, p_{\max}, hm), i = 0,1,..., p_{\max} - 1\right\} \tag{11}$$

As it was mentioned earlier, increasing the *h* value may cause oversmoothing of the regression function. It means, that some of the time series values can be underestimated. Underestimation α_i of the single value y_i can be expressed as the fraction of the estimation result $\hat{y}_i$ and the real value y_i:

$$\alpha_i = \frac{\hat{y}_i}{y_i} \tag{12}$$

and the underestimation on the tune set as a median of individual values:

$$\alpha = \text{med}\{\alpha_i, i = 1,..., k\} \tag{13}$$

where *k* is the number of objects in the test set.

Finally, the equation of the estimated value of the time series value at $t+p_{max}$, with α as an underestimation on the tune set and *m* as the number of points in the whole train set, can be presented as follows:

$$x_{t+p_{\max}} = \hat{f}(x_t) = \frac{1}{\alpha} \frac{\sum_{i=1}^{m-p_{\max}} x_{i+p_{\max}} K\left(\frac{x_i - x_t}{\mu h}\right)}{\sum_{i=1}^{m-p_{\max}} K\left(\frac{x_i - x_t}{\mu h}\right)} \tag{14}$$

3.2. *HKSVR* – local Support Vector Regression

Support Vector Regression is also the method, that can be applied to time series prediction [5, 19, 21, 24]. The author presents new local hybrid modification of *SVR*, called *HKSVR* (*Hybrid Kernel and Support Vector Regression*), that joins standard *SVR* and the definition of local *h*-neighborhood from *HASKE* model with its heuristic of *h* evaluation.

First step of the algorithm is the choice of the δ parameter value. It defines the way of transformation of time series into the modified space, in similar way as it was described in the beginning of the section 3. Its value can be evaluated after the Fourier analysis of the time series. Only periods of Fourier components with the highest values of their amplitude are considered as interesting values of the δ parameter. Then, the heuristic algorithm from *HASKE* evaluates the new value of *h* parameter for the train set with the usage of the μ parameter, as it was described in the previous section. Afterwards for every test point in the modified space its μh-neighborhood is treated as a train test for a local *SVR*. Finally, prediction of the $x_{t+\delta}$ time series value is equal to estimating of the regression function at the point x_t in the modified space, with the usage of support vectors.

4. Results

Performed experiments correspond to the rule of unbiased prediction, and the *ex post* error is a measure of the estimation accuracy. As the measure of the *ex post* error between the estimated value $\hat{y}_i$ and the real value y_i the *Root Mean Squared Error* (*RMSE*) is used:

$$err = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2} \tag{15}$$

where *k* is the number of object in the test set or the prediction horizon.

4.1. HASKE results

Four time series were used as the target of *HASKE* prediction ability: two of them were synthetic (*M* and *N* series) and two of them were taken from [4] (*G* and *E* series). Table 1 shows the comparison of prediction results from two models: Nadaraya—Watson estimator and *HASKE* model.

Table 1

Comparison of Nadaraya-Watson estimator and *HASKE* model

Time series	Period duration	NW	*HASKE*
M	17	42,77	8,74
N	8	49,09	4,60
G	12	275,26	17,18
E	11	33,37	36,76

It can be observed that for only one time series the HASKE prediction error was greater than the NW estimator error.

4.2. HKSVR results

The analyzed time series is the S&P500 stock index [1]. As a data sample closing values from 02.01.2003 to 18.12.2006 quotation were used. The value of δ parameter was evaluated as the local maximum in the discrete Fourier transform amplitude. For each harmonic its amplitude was calculated and consecutive harmonics with the highest amplitudes were chosen, excluding the global maximum. The δ for the first harmonic (the highest amplitude) correspond to the constant component so it was rejected from the further analysis.

The Table 2 shows differences between the errors as the result of using *SVR* and the *HKSVR*. It means that positive values speaks for the benefit of *HKSVR* (the increase of prediction accuracy).

Table 2

The decrease of estimation error after the usage of HKSVR model

Prediction horizon	n^{th} maximal harmonic in Fourier trasform				
	2	3	4	5	6
1	21,84	21,80	5,22	0,00	0,00
2	5,11	-9,81	-37,34	0,00	-2,26
3	2,55	10,22	-49,84	0,00	2,75
4	-0,67	-2,30	-46,29	0,00	1,67
5	-0,22	-4,19	-8,36	44,26	0,35
6	4,51	7,30	69,22	-0,62	19,31
7	5,80	18,46	-33,63	-1,08	-252,94
8	5,24	13,92	-35,22	-0,58	16,34
9	5,63	13,04	-25,30	-2,24	-4,52
10	8,99	16,89	-22,58	-2,80	-381,58
avg	**5,88**	**8,53**	**-18,41**	**3,69**	**-60,09**
std	**6,31**	**10,61**	**35,03**	**14,29**	**139,10**

Two last rows of the Table 2 describe the mean error decrease and the standard deviation of the error decrease. It is easy to notice, that in general only 2^{nd} and 3^{rd} harmonic give satisfying error improvement.

Let us consider the "rate of return" time series, defined in the following way:

$$r_x(t) = \frac{x(t) - x(t-1)}{x(t-1)} \tag{16}$$

That time series represents relative changes of time series value between two consecutive moments. It occurs, that the *HKSVR* model does not improve the result of the support vector regression method. That fact does not surprise if the correlation of the time series in the modified space is considered. The average correlation of the S&P500 time series in the modified space, if prediction horizon changes from 1 to 10 and considered harmonics are from 2^{nd} to 5^{th}, takes the value 0,906 with the standard deviation 0,016. The same coefficients for S&P500 rates of returns time series take the value 0,007 (average correlation) and 0,027 (standard deviation). The very low correlation can be admitted as the reason, why the *HASKE* model does not improve *SVR* algorithm.

5. Conclusions

This short article describes two new kernel estimators of the regression function, that are results of the time series prediction research [20]. First of them (*HASKE*) helps to avoid the situation where there are no train objects for test objects. Its main advantage (improvement of an certainty that *h*-neighborhoods of test points will not be empty sets) is confirmed by results on synthetic and real time series.

Definition of the point neighborhood as the *h*-neighborhoods led the author to the local modification of *SVR* method, *HKSVR*. Results of the prediction of the S&P500 closing values were better for the *HKSVR* model of regression than for the *SVM* model. Author also points correlation as the simple criterion whether the *HKSVR* model may improve the typical global *SVR* algorithm.

BIBLIOGRAPHY

1. S&P500 historical data. http://stooq.pl/q/d/?s=s%26p500.
2. de Boor C.: A practical guide to splines. Springer, 2001.
3. Boser B. E., Guyon I. M., Vapnik V. N.: A training algorithm for optimal margin classifiers. In Proc. of the 5th annual workshop on Computational learning theory, Pittsburgh 1992, s. 144÷152.
4. Box G. E. P., Jenkins G. M.: Analiza szeregów czasowych. PWN, Warszawa 1983.
5. Cao L. J., Tay F. E. H.: Svm with adaptive parameters in financial time series forecasting. IEEE Trans. on Neural Networks, 14(6), 2003, s. 1506÷1518.

6. Cleveland W. S., Devlin S. J.: Locally weighted regression. Jour. of the Am. Stat. Ass., 83(403), 1988, s. 596÷610.
7. Epanechnikov V. A.: Nonparametric estimation of a multivariate probability density. Theory of Probability and Its Applications, 14, 1969, s. 153÷158.
8. Fan J., Gijbels I.: Variable bandwidth and local linear regression smoothers. Annals of Statistics, 20(4), 1992, s. 2008÷2036.
9. Fernandez R.: Predicting time series with a local support vector regression machine. In Proc. of the ECCAI Advanced Course on Artificial Intelligence '99.
10. Friedman J. H.: Multivariate adaptive regression splines. Annals of Statistics, 19(1), 1991, s. 1÷141.
11. Gajek L., Kałuszka M.: Wnioskowanie statystyczne. WNT, Warszawa 2000.
12. Gasser T., Kneip A., Kohler W.: A flexible and fast method for automatic smoothing. Jour. of the Am. Stat. Ass., 86(415), 1991, s. 643÷652.
13. Gasser T., Muller H. G.: Estimating regression function and their derivatives by the kernel method. Scandinavian Journal of Statistics, 11, 1984, s. 171÷185.
14. Hastie T. J., Tibshirani R. J.: Generalized Additive Models. Chapman & Hall/CRC, 1990.
15. Huang K., Yang H., King I., Lyu M.: Local svr for financial time series prediction. In Proc of IJCNN'06, Vancouver 2006, s. 1622÷1627.
16. Kaastra I., Boyd M.: Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10(3), 1996, s. 215÷236.
17. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT, Warszawa 2005.
18. Kulczycki P.: Estymatory jądrowe w analizie systemowej. WNT, Warszawa 2005.
19. Michalak M.: Możliwości poprawy jakości usług w transporcie miejskim poprzez monitoring natężenia potoków pasażerskich. In ITS dla Śląska, Katowice 2008.
20. Michalak M., Stąpor K.: Estymacja jądrowa w predykcji szeregów czasowych. Studia Informatica, Vol. 29, No 3A(78), Gliwice 2008, s. 71÷90.
21. Muller K. R., Smola A. J., Ratsch G., Scholkopf B., Kohlmorgen J., Vapnik V.: Predicting time series with support vector machines. In Proceedings of the 7th ICANN, LNCS(1327), Springer-Verlag, London 1997, s. 999÷1004.
22. Nadaraya E. A.: On estimating regression. Theory of Probability and Its Applications, 9(1), 1964, s. 141÷142.
23. Scholkopf B., Smola A.: Learning with Kernels. MIT Press, 2002.
24. Sikora M., Kozielski M., Michalak M.: Innowacyjne narzędzia informatyczne analizy danych. Wydział Transportu, Gliwice 2008.
25. Silverman B. W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, 1986.

26. Smola A. J.: Regression estimation with support vector learning machines. Master's thesis, Technische Universitat München 1996.
27. Smola A. J., Scholkopf B.: A tutorial on support vector regression. Statistics and Computing, 14(3), 2004, s. 199÷222.
28. Taylor J. S., Cristianini N.: Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
29. Terrell G. R.. The maximal smoothing principle in density estimation. Jour. of the Am. Stat. Ass., 85(410), 1990, s. 470÷477.
30. Terrell G. R., Scott D. W.: Variable kernel density estimation. Annals of Statistics, 20(3), 1992, s. 1236÷1265.
31. Turlach B. A.: Bandwidth selection in kernel density estimation: A review. Technical report, Universite Catholique de Louvain, 1993.
32. Vapnik V. N.: Statistical Learning Theory. Wiley, 1988.
33. Wand M. P., Jones M. C.: Kernel Smoothing. Chapman & Hall, 1995.
34. Watson G. S.: Smooth regression analysis. Sankhya - The Indian Journal of Statistics, 26(4), 1964, s. 359÷372.

Recenzent: Prof. dr hab. inż. Jacek Koronacki

Wpłynęło do Redakcji 15 stycznia 2009 r.

Omówienie

Artykuł podejmuje temat nieparametrycznej predykcji szeregów czasowych. Problem ten wiąże się z zastosowaniem metod nieparametrycznej regresji do zadania predykcji. Jednym z takich rozwiązań jest zastosowanie estymatora Nadarayi-Watsona w zmodyfikowanej przestrzeni cech oraz przy użyciu adaptacyjnej metody wyznaczania wartości parametru wygładzającego [20].

W artykule opisano dwa nowe algorytmy, nawiązujące do estymatorów jądrowych i maszyny wektorów podpierających. Główną zaletą pierwszego z nich (*HASKE*) jest to, że w znaczący sposób eliminuje problem pustych zbiorów uczących, będących otoczeniem punktu testowego. Wynika to z adaptacyjnej metody wyznaczania wartości parametru wygładzającego h, który determinuje zbiór uczący dla punktów testowych. Aby z kolei zapobiec zbytniemu wygładzeniu danych, wyznacza się adaptacyjnie inny współczynnik, zwany niedoszacowaniem.

Drugi algorytm łączy w sobie *SVR* i *HASKE*. Tak powstały model (*HKSVR*) można potraktować jako lokalną modyfikację *SVR*. Dla każdego punktu uczącego wyznacza się jego *h*-otoczenie na podstawie strategii algorytmu *HASKE*. Następnie tak uzyskany podzbiór zbioru uczącego staje się zbiorem uczącym dla algorytmu *SVR*.

Oba algorytmy zostały zastosowane dla predykcji finansowych szeregów czasowych. Okazuje się jednak, że model HKSVR nie zawsze jest w stanie poprawić wyniki regresji za pomocą SVR. Jednym z kryteriów, wskazujących na celowość użycia modelu *HKSVR*, jest analiza korelacji szeregu czasowego w zmodyfikowanej przestrzeni cech.

Address

Marcin MICHALAK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, Marcin.Michalak@polsl.pl