СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2015. – № 3(81).** – 69–78

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.23

ПОСТРОЕНИЕ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КВАДРАТИЧНОЙ ФУНКЦИИ ПОТЕРЬ В МЕТОДЕ ОПОРНЫХ ВЕКТОРОВ*

А.А. ПОПОВ[1], Ш.А. БОБОЕВ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, профессор кафедры теоретической и прикладной информатики. E-mail: a.popov@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант кафедры теоретической и прикладной информатики. E-mail: shboboev@mail.ru*

Рассматривается один из методов непараметрического оценивания регрессионных зависимостей, принадлежащий к классу ядерных методов, – метод опорных векторов с квадратичной функцией потерь, который является модификацией алгоритма опорных векторов. Приводится исходная задача оптимизации для получения параметров регрессионной модели. Получаемое в явном виде решение выписывается в терминах двойственных переменных. То, что решение удается получить в явном виде, выгодно отличает LS SVM от базового SVM, требующего решения квадратичной задачи с ограничениями. В данной работе для LS SVM использовались полиномиальное и гауссово ядро. Проводится исследование возможности использования критерия скользящего контроля и критерия регулярности для настройки внутренних параметров алгоритма опорных векторов с квадратичной функцией потерь. Точность получаемых решений контролируется с использованием среднеквадратичной ошибки. Вычислительный эксперимент проводился на модельных данных. В качестве модели, порождающей данные, была выбрана нелинейная зависимость от входного фактора. Дисперсия помехи (уровень шума) определялась в процентах от мощности незашумленного сигнала. В работе в табличной форме отражены результаты восстановления зависимости с использованием гауссового ядра при фиксированном значении параметра регуляризации. Качество решения иллюстрируется графически. В работе сравнивались также качество восстановления зависимостей при использовании полиномиального или гауссова ядра. В работе делается вывод, что для настройки внутренних параметров алгоритма LS SVM можно использовать критерий скользящего контроля и критерий регулярности.

Ключевые слова: регрессия, метод опорных векторов, квадратичная функция потерь, критерий скользящего контроля, критерий регулярности, коэффициент регуляризации, ядерная функция, среднеквадратичная ошибка, полиномиальное ядро, RBF-ядро

DOI: 10.17212/2307-6879-2015-3-69-78

---

* Статья получена 18 мая 2015 г.

ВВЕДЕНИЕ

Алгоритм опорных векторов с квадратичной функцией потерь (LS-SVM) [1] как с линейными, так и с нелинейными ядерными функциями – один из наиболее перспективных, основанных на обучении алгоритмов построения регрессии. Он является модификацией алгоритма опорных векторов (SVM) с функцией нечувствительности Вапника [2–7]. Одним из важных этапов построения регрессии с использованием метода опорных векторов является настройка его ряда внутренних параметров. При использовании произвольных значений параметров алгоритма опорных векторов качество работы алгоритма может существенно варьироваться. Эвристический подход по априорному выбору внутренних параметров SVM-алгоритма предлагается в работах [8, 9]. Некоторые вопросы настройки параметров с использованием вложенных сеток рассмотрены в работах [13, 15].

Важным моментом в решении задачи настройки параметров алгоритма опорных векторов является выбор критерия качества получаемых решений. Известным и активно развиваемым подходом для выбора линейных параметрических моделей оптимальной сложности является использование так называемых внешних критериев [10–12, 14]. В нашем случае в качестве таковых могут быть использованы различные варианты критерия скользящего контроля и регулярности. В данной работе исследуется возможность критериев этого типа быть использованными при настройке параметров алгоритма опорных векторов. Для проведения вычислительных экспериментов разработан программный модуль, реализующий LS SVM с настройкой его параметров.

1. LS-SVM РЕГРЕССИЯ

Рассмотрим задачу восстановления зависимости по зашумленным данным. Дана обучающая выборка $D_n = \{(x_k, y_k) : x_k \in X, y_k \in Y; k = 1, \ldots, n\}$ объема n наблюдений вида

$$y_k = m(x_k) + e_k, k = 1, ..., n, \quad (1)$$

где $e_k \in R$ будем считать независимо и одинаково распределенной ошибкой с $E[e_k \mid X = x_k] = 0$ и $Var[e_k] = \sigma^2 < \infty$, $m(x)$ – неизвестная действительная гладкая функция и $E[y_k \mid x = x_k] = m(x_k)$. Нашей целью является поиск пара-

метров ω и b исходного пространства, которые минимизируют эмпирический функционал риска

$$R_{emp}(\omega,\ b) = \frac{1}{n}\sum_{k=1}^{n}\left(\left(\omega^T\varphi(x_k)+b\right)-y_k\right)^2. \tag{2}$$

Задачу нахождения вектора ω и $b \in R$ можно свести к решению следующей задачи оптимизации [1]:

$$\min_{\omega,b,e} J(\omega,\ e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{n}e_k^2 \tag{3}$$

в предположении, что $y_k = \omega^T\varphi(x_k)+b+e_k,\ k=1,\ldots,n$.

Решение задачи (3) обычно проводят в двойственном пространстве с использованием функционала Лагранжа

$$L(\omega,\ b,\ e,\ \alpha) = J(\omega,\ e) - \sum_{k=1}^{n}\alpha_k\left(\omega^T\varphi(x_k)+b+e_k-y_k\right) \tag{4}$$

с лагранжевыми множителями $\alpha_k \in R$.

Условия оптимальности задаются следующим образом:

$$\begin{cases} \dfrac{dL}{d\omega}=0\rightarrow\omega=\sum\limits_{k=1}^{n}\alpha_k\varphi(x_k);\quad \dfrac{dL}{de_k}=0\ \rightarrow\alpha_k=\gamma e_k,\ k=1,\ ...,\ n; \\ \dfrac{dL}{db}=0\rightarrow\sum\limits_{k=1}^{n}\alpha_k=0;\quad \dfrac{dL}{d\alpha_k}=0\ \rightarrow\omega^T\varphi(x_k)+b+e_k=y_k,\ k=1,\ ...,\ n. \end{cases} \tag{5}$$

После исключения ω и e получаем решение

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega+\dfrac{1}{\gamma}I_n \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{6}$$

где $y=(y_1,\ldots,y_n)^T$, $1_n=(1,\ldots,1)^T$, $\alpha=(\alpha_1,\ldots,\alpha_n)$ и $\Omega_{kl}=\varphi(x_k)^T\varphi(x_l)$ для $k,\ l=1,\ldots,n$. Результирующая LS-SVM модель имеет вид

$$y_n(x) = \sum_{k=1}^{n}\alpha_k K(x,\ x_k)+\hat{b}, \tag{7}$$

где $K(x, x_k)$ – ядро скалярного произведения,

$$\hat{b} = \frac{1_n^T \left(\Omega + \frac{1}{\gamma} I_n \right)^{-1} y}{1_n^T \left(\Omega + \frac{1}{\gamma} I_n \right)^{-1} 1_n}, \quad \alpha = \left(\Omega + \frac{1}{\gamma} I_n \right)^{-1} \left(y - 1_n \hat{b} \right). \tag{8}$$

В данной работе для оценки качества разработанного алгоритма используются MSE (*Mean Square Error*), *критерий скользящего контроля* LOO CV (*Leave-One-Out Cross Validation*) и *критерий регулярности* K-FOLD CV (*K-Fold Cross Validation*).

Контроль по отдельным объектам (leave-one-out, LOO*)* является частным случаем полного скользящего контроля. Преимущества LOO в том, что каждый объект ровно один раз участвует в контроле, а длина обучающих подвыборок лишь на единицу меньше длины полной выборки. Недостатком LOO является большая ресурсоемкость, поскольку задачу обучения приходится решать l раз, что сопряжено со значительными вычислительными затратами.

Критерий регулярности K-FOLD CV состоит в том, что исходная выборка разбивается некоторое количество раз на обучающую и контрольную объемом в *K* наблюдений с усреднением результатов.

2. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Целью вычислительного эксперимента являлось выяснение возможности выбирать внутренние параметры алгоритма, а это параметр регуляризации γ и параметры ядерных функций, с учетом значений критериев LOO и K-FOLD CV. Для проведения исследования использовалась следующая тестовая функция: $y = 7 \Big/ \left(e^{(x+0,75)^2} + 3x \right)$. Вид моделируемой помехи – гауссовский шум. В качестве ядер использовались полиномиальное и гауссово. Уровень помехи выбирался как 5 %, 10 % и 20 % от мощности незашумленного сигнала. Количество наблюдений выбиралось равным 100.

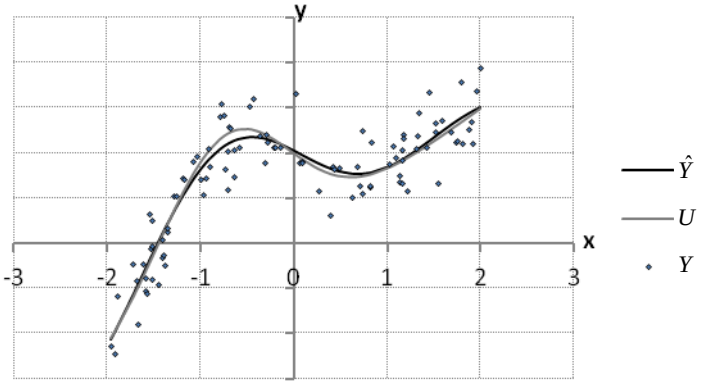
Решения, полученные с использованием полиномиального ядра, оказались несколько хуже по значению MSE. Ниже приведены результаты выбора оптимального значения масштаба σ^2 для гауссова ядра при зафиксированном значении параметра регуляризации $\gamma = 500$.

Значения MSE, LOO CV и K-FOLD при уровне шума 20 %

σ^2	MSE	LOO CV	K-FOLD
0,1	0,09561587400	0,00053752700	2,39671539600
0,12589254	0,08211723800	0,00052625000	2,32836776100
0,15848932	0,06576996500	0,00051941900	2,28432100300
0,19952623	0,05135821400	0,00051680600	2,26494979800
0,25118864	0,04035187200	0,00051528000	2,24461674400
0,31622777	0,03606270700	0,00051351400	2,22030423900
0,39810717	0,03436424500	0,00050779800	2,18764072200
0,50118723	0,03153968000	0,00050067500	2,15999533100
0,63095734	0,02847867500	0,00049467100	2,14364818500
0,79432824	0,02704932100	0,00048992400	2,13201871000
1	0,02730739600	0,00048529100	2,11730118900
1,25892541	0,02924289200	0,00048040300	2,09809860600
1,58489319	0,03429909600	0,00047647300	2,07693503600
1,99526232	0,04669731700	**0,00047535600**	**2,05976715400**
2,51188643	0,07252887400	0,00047939700	2,06091960400
3,16227766	0,11096565000	0,00048808600	2,08558729700
3,98107171	0,15853652800	0,00049948900	2,13536721500
5,01187234	0,22111258500	0,00051548700	2,23179551000
6,30957345	0,32073038900	0,00054458200	2,42873128700
7,94328235	0,49412138000	0,00060160800	2,80479643400
10	0,76676804300	0,00069769600	3,39834513000

Критерий K-FOLD вычислялся по схеме: усредненное по двадцати испытаниям значение ошибки прогноза в точке случайно выбираемой тестовой части выборки объемом в 10 наблюдений. Из таблицы видно, что использование критериев качества как LOO CV так и K-FOLD позволяет выбрать параметр масштаба гауссова ядра близким к тому, что выбирается на основе среднеквадратичной ошибки.

Качество восстановленной зависимости при выбираемой $\sigma^2 = 1,99526232$ иллюстрируется на рисунке.



Графики зависимостей:

U – незашумленный отклик, Y – зашумленный отклик,
$\hat{Y}$ – решение по построенной модели

ЗАКЛЮЧЕНИЕ

В рамках проведенных исследований LS-SVM показал себя как достаточно устойчивый метод восстановления зависимостей по зашумленным данным. Даже при относительно высоком уровне шума модель, получаемая с использованием гауссова ядра обладает хорошей обобщающей способностью. Важную роль в точности получаемых решений играет правильный выбор параметров ядерных функций и коэффициента регуляризации. Выбор этих параметров можно проводить, опираясь на критерии скользящего контроля типа LOO CV и K-FOLD. При этом перебор решений целесообразно проводить на сетке оптимизируемых параметров или с использованием вложенных сеток, как в работе [3]. Выбор типа используемого ядра необходимо проводить через вариативное моделирование.

СПИСОК ЛИТЕРАТУРЫ

1. Least squares support vector machines / J.A.K. Suykens, T. van Gestel, J. de Brabanter, B. de Moor, J. Vandewalle. – New Jersey; London; Singapore; Hong Kong: World Scientific, 2002. – 290 p.

2. *Smola A.* Regression estimation with support vector learning machines: master's thesis. – Munchen: Technische Universitat, 1996. – 78 p.

3. *Smola A.* Learning with kernels: PhD thesis in computer science. – Berlin: Technische Universitat, 1998. – 210 p.

4. *Smola A.* A tutorial on support vector regression // Statistics and Computing. – 2004. – N 14. – P. 199–222.

5. *Vapnik V.N.* Estimation of dependences based on empirical data. – New York: Springer Verlag, 1982. – 399 p.

6. *Vapnik V.* Statistical learning theory. – New York: John Wiley, 1998. – 736 p.

7. *Vapnik V.N.* The nature of statistical learning theory. – New York: Springer Verlag, 1995. – 188 p.

8. C*herkassky V., Ma Y.Q.* Practical selection of SVM parameters and noise estimation for SVM regression // Neural Networks. – 2004. – N 17. – P. 113–126.

9. *Huang C.M., Lee Y.J.* Model selection for support vector machines via uniform design // Computational Statistics & Data Analysis. – 2007. – N 52. – P. 335–346.

10. *Лисицин Д.В., Попов А.А.* Конструирование критериев селекции многомерных регрессионных моделей // Сборник научных трудов НГТУ. – 1996. – № 1. – С. 13–20.

11. *Попов А.А.* Планирование эксперимента в задачах структурного моделирования с использованием критерия скользящего прогноза // Заводская лаборатория. – 1996. – № 10. – С. 42–44.

12. *Попов А.А.* Разбиение выборки для внешних критериев селекции моделей с использованием методов планирования эксперимента // Заводская лаборатория. – 1997. – № 1. – С. 49–53.

13. Попов А.А, Саутин А.С. Определение параметров алгоритма опорных векторов при решении задачи построения регрессии // Сборник научных трудов НГТУ. – 2008. – № 2 (52). – С. 35–40.

14. *Попов А.А.* Оптимальное планирование эксперимента в задачах структурной и параметрической идентификации моделей многофакторных систем: монография. – Новосибирск: Изд-во НГТУ, 2013. – 296 с.

15. Popov A.A. Sautin A.S. Selection of support vector machines parameters for regression using nested grids // The third international forum on strategic technology (IFOST 2008): proceedings, Novosibirsk–Tomsk, Russia, 23–29 June 2008. – Novosibirsk, 2008. – P. 329–331.

Попов Александр Александрович, доктор технических наук, профессор кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных иссле-

дований – методы анализа данных, оптимальное планирование экспериментов. Имеет более 150 научных работ, в том числе 2 монографии. E-mail: a.popov@corp.nstu.ru

Бобоев Шараф Асрорович, аспирант кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. E-mail: shboboev@mail.ru

The Construction of a Regression Relationships using Least Square in Support Vector Machines*

A.A. Popov[1], Sh.A. Boboev[2]

[1.] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, D. Sc. (Eng.), professor. E-mail: a.popov@corp.nstu.ru*
[2.] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, post-graduate student. E-mail: shboboev@mail.ru*

Considered one of the methods of nonparametric estimation of regression dependencies belonging to the class of kernel methods – support vector machines with least squares (LS-SVM), which is a modification of the algorithm of support vector machine (SVM). Conducted the original optimization problem to obtain the parameters of the regression model. The resulting explicit solution is written in terms of dual variables. That the solution is possible to obtain in explicit form it distinguishes LS-SVM from basic SVM, requiring the solution of the quadratic problem with constraints. In this paper, for support vector machines with a least squares was used the polynomial and the Gaussian kernel. Conduct research of the possibility of using cross-validation criteria for tuning the internal parameters of support vector machine algorithm with a least square. The accuracy of the solutions is controlled by using the mean squared error. A computational experiment was performed on simulated data. As a model generating the data was selected nonlinear dependence on the input factor. The variance of the noise (noise level) was determined as a percentage of the power no-noisy signal. In the work in tabular form is reflected the results of the recovery dependence using a Gaussian kernel when a fixed value of the regularization parameter. The quality of decisions is illustrated graphically. In paper compared the quality of the recovery of dependencies when using polynomial and Gaussian kernel. In work concludes that to configure the internal parameters of the algorithm support vector machines with least squares it is possible to use the criterion of cross-validation and criterion of regularity.

Keywords: regression, support vector machines, least square, criterion of a cross-validation, criterion of a regularity, factor of regularization, kernel function, mean square error, polynomial kernel, RBF kernel

DOI: 10.17212/2307-6879-2015-3-69-78

* *Received 18 May 2015.*

REFERENCES

1. Suykens J.A.K., Gestel T. van, Brabanter J. de, Moor B. de, Vandewalle J. *Least square support vector machines*. New Jersey, London, Singapore, Hong Kong, World Scientific. 290 p.

2. Smola A. *Regression estimation with support vector learning machines: master's thesis*. München, Technische Universitat, 1996. 78 p.

3. Smola A. *Learning with kernels*. PhD thesis in computer science. Berlin, Technische Universitat, 1998. 210 p.

4. Smola A. A tutorial on support vector regression. *Statistics and Computing*, 2004, no. 14, pp. 199–222.

5. Vapnik V.N. Estimation of dependences based on empirical data. New York, Springer Verlag, 1982. 399 p.

6. Vapnik V. *Statistical learning theory*. New York, John Wiley, 1998. 736 p.

7. Vapnik V.N. *The nature of statistical learning theory*. New York, Springer Verlag, 1995. 188 p.

8. Cherkassky V., Ma Y.Q. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 2004, no. 17, pp. 113–126.

9. Huang C.M., Lee Y.J. Model selection for support vector machines via uniform design. *Computational Statistics & Data Analysis*, 2007, no. 52, pp. 335–346.

10. Lisitsin D.V., Popov A.A. Konstruirovanie kriteriev selektsii mnogomernykh regressionnykh modelei [Construction of the breeding criteria of multivariate regression models]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1996, no. 1, pp. 13–20.

11. Popov A.A. Planirovanie eksperimenta v zadachakh strukturnogo modelirovaniya s ispol'zovaniem kriteriya skol'zyashchego prognoza [Planning of experiment in problems of structural modeling using the criterion of rolling forecasts]. *Zavodskaya laboratoriya – Factory Laboratory*, 1996, no. 10, pp. 42–44.

12. Popov A.A. Razbienie vyborki dlya vneshnikh kriteriev selektsii modelei s ispol'zovaniem metodov planirovaniya eksperimenta [Splitting the sample for external criteria of selection models using methods of experiment planning]. *Zavodskaya laboratoriya – Factory Laboratory*, 1997, no. 1, pp. 49–53.

13. Popov A.A, Sautin A.S. Opredelenie parametrov algoritma opornykh vektorov pri reshenii zadachi postroeniya regressii [Determination of parameters of support vector machine algorithm when solving the problem of constructing the regression]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 2008, no. 2 (52), pp. 35–40.

14. Popov A.A. *Optimal'noe planirovanie eksperimenta v zadachakh strukturnoi i parametricheskoi identifikatsii modelei mnogofaktornykh sistem* [The optimal planning of experiment in problems of structural and parametric identification of models of multifactor systems: monograph]. Novosibirsk, NSTU Publ., 2013. 296 p.

15. Popov A.A. Sautin A.S. Selection of support vector machines parameters for regression using nested grids. *The third international forum on strategic technology (IFOST 2008): proceedings*, Novosibirsk–Tomsk, Russia, 23–29 June 2008, pp. 329–331.