ISSN 1814-1196
Научный вестник НГТУ
том 65, № 4, 2016, с. 80–99

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 65, No. 4, 2016, pp. 80–99

ОБРАБОТКА ИНФОРМАЦИИ INFORMATION PROCESSING

УДК 519.23

Получение тестовой выборки в методе LS–SVM с использованием оптимального планирования эксперимента*

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. E-mail: a.popov@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант. E-mail: shboboev@mail.ru*

В работе рассматривается задача восстановления регрессионной зависимости по методу опорных векторов с квадратичной функцией потерь. Данный метод относится к классу ядерных методов. Для настройки ряда внутренних параметров алгоритма LS–SVM обсуждается проблема получения тестовой выборки. Приведены различные критерии селекции моделей, которые основываются на разбиении выборки на обучающую и тестовую части. Проблема разбиения выборки на тестовую и обучающую части с использованием метода *D*-оптимального планирования эксперимента подробно рассмотрена для случая линейных параметрических регрессионных моделей. Данный метод получения тестовой выборки предложено использовать для метода LS–SVM. Приводится последовательный алгоритм получения обучающей и тестовой частей выборки наблюдений применительно к методу LS–SVM. Для проверки работоспособности предлагаемого метода разбиения выборки проведен вычислительный эксперимент. В нем повышение точности решений по LS–SVM проводилось посредством подбора масштаба гауссовой ядерной функции. Данный параметр ядерной функции подбирался по минимуму ошибки прогноза на тестовой части выборки. Окончательно точность получаемых решений проверялась по среднеквадратичной ошибке. 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 сигнала. Сравнивались два способа разбиения выборки на обучающую и тестовую: случайное разбиение и разбиение по методу *D*-оптимального планирования эксперимента. Для выбора параметров алгоритма LS–SVM использовался также критерий перекрестной проверки. Результаты проведенных вычислительных экспериментов приведены в отдельных таблицах и рисунках. По результатам проведенных вычислительных экспериментов делаются выводы о том, что эффективность использования случайной тестовой выборки нестабильна и во много определяется конкретным вариантом разбиения. При этом стабильность результатов использования тестовой выборки, полученной при *D*-оптимальном разбиении выборки, значительно выше.

Ключевые слова: регрессия, метод LS–SVM, квадратичная функция потерь, тестовая выборка, обучающая выборка, оптимальное планирование эксперимента, *D*-оптимальный

---

* *Статья получена 30 августа 2016 г.*

план, критерий регулярности, критерий скользящего контроля, коэффициент регуляризации, ядерная функция, среднеквадратичная ошибка

DOI: 10.17212/1814-1196-2016-4-80-99

ВВЕДЕНИЕ

Метод опорных векторов с квадратичной функцией потерь (LS–SVM) как с линейными, так и с нелинейными ядерными функциями – один из наиболее перспективных алгоритмов построения регрессии. Он является модификацией алгоритма опорных векторов (SVM) с функцией нечувствительности Вапника. Одним из важных этапов построения регрессии с использованием метода опорных векторов является настройка ряда его внутренних параметров. При использовании произвольных значений параметров алгоритма опорных векторов качество работы алгоритма может существенно варьироваться.

Ключевым 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. В нашем случае в качестве таковых могут быть использованы различные варианты критериев, связанных с точностью прогноза на тестовой выборке. В данной работе исследуется возможность разбиения выборки на обучающую и тестовую части с привлечением методов оптимального планирования эксперимента.

ВНЕШНИЕ КРИТЕРИИ СЕЛЕКЦИИ МОДЕЛЕЙ

При построении моделей, описывающих поведение отклика от действующих факторов, главной задачей является определение структуры модели, поскольку, как правило, она априори неизвестна. Исследователь сталкивается с проблемой выбора структуры модели. Для решения этой задачи назначаются определенные критерии «качества», которым должна удовлетворять искомая модель. Будем в дальнейшем называть их критериями селекции моделей. Перечень используемых критериев селекции достаточно широк и подробно представлен в обзорах [1–4].

Критерии селекции моделей можно поделить на две группы: критерии, использующие всю выборку данных, и критерии, основанные на разбиении выборки на части.

Критерии, основанные на разбиении выборки на части

Пусть модель объекта подчиняется следующему уравнению наблюдения:

$$Y = \widehat{Y} + \varepsilon = \widehat{X}\widehat{\beta} + \varepsilon, \tag{1}$$

где $\widehat{Y} - (n \times 1)$ – вектор ненаблюдаемого незашумленного выхода объекта; $\widehat{X} - (n \times m)$ – расширенная матрица плана, соответствующая истинному набору регрессоров $\hat{x}_1,...,\hat{x}_m$; $\varepsilon-(n\times 1)$ – вектор ненаблюдаемых случайных ошибок измерения, относительно которых выполнены предположения $E(\varepsilon)=0_n$, $E(\varepsilon\varepsilon^T)=\sigma^2 I_n$, где 0_n – вектор, состоящий из нулей; σ^2 – неизвестная дисперсия наблюдения; I_n – единичная матрица размера n. Набор регрессоров $\hat{x}_1,...,\hat{x}_m$ образует множество $\hat{X}$, о котором известно, что $\hat{X}\subset\Re$, где $\Re$ – некоторое расширенное множество регрессоров. Пусть в результате наблюдения объекта получена $Z-(n\times p)$ – расширенная матрица плана из n наблюдений над p регрессорами из $\Re$, и требуется определить множество $\hat{X}$ и получить оценку параметров $\hat{\beta}$. Для поиска наилучшей аппроксимации для (1) воспользуемся каким-либо переборным алгоритмом. Пусть $X-(n\times s)$ – расширенная матрица наблюдений для текущей модели из s регрессоров, образующих множество $L\subset\Re$. Регрессия отклика y по L будет определяться по уравнению наблюдения

$$y = X\theta + e, \tag{2}$$

где $e-(n\times 1)$ – вектор ненаблюдаемых случайных ошибок измерения, относительно которых выполнены предположения $E(e)=0_n$, $E(ee^T)=\sigma^2 I_n$.

Предположим, что выборка наблюдений W разбита на две части *A* и *B*. В методах структурной оптимизации активно используются следующие так называемые внешние критерии селекции моделей [4–6]:

- *критерий регулярности*

$$\Delta^2(B)=\Delta^2(B/A)=\left\|y_B - X_B\hat{\theta}_A\right\|^2,$$

где запись $\Delta^2(B/A)$ означает «ошибка» на выборке *B* модели, коэффициенты которой получены с использованием выборки *A*;

- *критерий симметричной регулярности*

$$d^2=\Delta^2(B/A)+\Delta^2(A/B)=\left\|y_B - X_B\hat{\theta}_A\right\|^2+\left\|y_A - X_A\hat{\theta}_B\right\|^2;$$

- *критерий стабильности*

$$S^2=\Delta^2(A\cup B/A)+\Delta^2(A\cup B/B)=\left\|y_W - X_W\hat{\theta}_A\right\|^2+\left\|y_W - X_W\hat{\theta}_B\right\|^2;$$

- *критерий непротиворечивости*

$$n_{CM}^2=\left\|X_W\hat{\theta}_A - X_W\hat{\theta}_B\right\|^2;$$

- *критерий несмещенности по коэффициентам*

$$n_C^2=\left\|\hat{\theta}_A-\hat{\theta}_B\right\|^2;$$

• *критерий вариативности*

$$V^2 = \left(X_W \hat{\theta}_A - X_W \hat{\theta}_W \right)^T \left(X_W \hat{\theta}_W - X_W \hat{\theta}_B \right).$$

К рассматриваемой группе критериев относится также критерий «скользящего контроля» (CV – cross validation)

$$\Delta_{ck}^2 = \sum_i (y_i - f^T(x_i)\hat{\theta}_{(i)})^2 ,$$

где $\hat{\theta}_{(i)}$ – оценка параметров по выборке W с исключенным i-м наблюдением.

Теоретическое обоснование внешних критериев проведено в работах [5–9]. Анализ этих работ показывает, что в отношении методов структурной идентификации складывается теория, в основу которой положен принцип J-оптимальности модели. Рассмотрим его.

J-оптимальная модель определяется решением задачи

$$f^* = Arg \min_{f \in \Omega_f} J(f), \tag{3}$$

где Ω_f – множество всех возможных моделей, формируемых на основе наблюдаемой матрицы Z. Теоретическая (идеальная) модель $J(f)$ определяет собой среднеквадратичную ошибку предсказания истинного отклика либо на всей выборке, либо на прогнозной части B:

$$J(f) = \frac{1}{n} E \left\| \hat{y} - X\hat{\theta} \right\|^2, \quad J_B(f) = \frac{1}{n_B} E \left\| \hat{y}_B - X_B \hat{\theta}_A \right\|^2 .$$

При решении задачи (3) минимуму $J(f)$ соответствует оптимальная сглаживающая модель, а минимуму $J_B(f)$ – оптимальная прогнозирующая. Теоретическое исследование критериев $J(f)$ и $J_B(f)$ показало, что в условиях шума с нулевой дисперсией минимумы этих критериев приходятся на модель сложности $s^0 = m$. При дисперсии наблюдения $\sigma^2 > 0$ функции $J(f)$ и $J_B(f)$ имеют единственный минимум в точке $s^* \le s^0$. С ростом σ^2 сложность s^* уменьшается, т. е. J-оптимальной становится все более простая модель. В пределе при относительно сильной зашумленности данных в качестве модели оптимальной сложности будет выбираться модель среднего. В качестве оценок для идеальных критериев $J(f)$ и $J_B(f)$ могут выступать внешние критерии $\Delta^2(B)$, d^2, S^2, n_{CM}^2, Δ_{ck}^2. Исследования [5–9] показывают, что внешние критерии $\Delta^2(B)$, d^2, S^2, Δ_{ck}^2 несмещенно оценивают J-оптимальную модель. Критерий n_{CM}^2 при определенных условиях также несмещенно оценивает J-оптимальную модель. В то же время критерий в виде остаточной суммы квадратов RSS, как и его скорректированная величина $RSS/(n-s)$, не обладает необходимыми свойствами помехоустойчивости, поскольку минимум этих критериев при $\sigma^2 > 0$ не находится в области $s \le s^0$.

Предлагаемые к использованию внешние критерии допускают достаточно простую статистическую интерпретацию с позиций проверки линейных гипотез [10]. Предположим, что ошибки наблюдения распределены по нормальному закону $N\left(0,\ \sigma^2 I_N\right)$. Пусть априорные данные о параметрах модели определены информацией из выборки A и состоят в выполнении $X_A^T X_A \theta = X_A^T y_A$. Поступают новые данные в виде выборки B. По выборке B производится оценивание параметров и осуществляется проверка гипотезы $H: X_A^T X_A \theta = X_A^T y_A$. Обозначим $RSS = (y_B - X_B \hat{\theta}_B)^T (y_B - X_B \hat{\theta}_B)$, $RSS_H = (y_B - X_B \hat{\theta}_H)^T (y_B - X_B \hat{\theta}_H)$, где $\hat{\theta}_H$ – оценки, полученные при ограничениях $X_A^T X_A \theta = X_A^T y_A$. В нашем случае при условии полного ранга матрицы $X_A^T X_A$ имеем, что $\hat{\theta}_H = \hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A$; F-статистика для проверки гипотезы будет иметь вид

$$F = \frac{(RSS_H - RSS)/s}{RSS/(n_B - s)} = \frac{(\Delta^2(B) - \varepsilon^2(B))/s}{\varepsilon^2(B)/(n_B - s)}.$$

С учетом взаимосвязи критериев числитель F-статистики можно записать в виде $RSS_H - RSS = n_{CM}^2(B) = \left(\hat{\theta}_B - \hat{\theta}_A\right)^T X_B^T X_B \left(\hat{\theta}_B - \hat{\theta}_A\right)$. Принятие гипотезы зависит, таким образом, от нормированных $n_{CM}^2(B)$, $\varepsilon^2(B)$. Изменим теперь постановку задачи. Пусть теперь параметры оцениваются на всей выборке $W = A \bigcup B$, и по-прежнему проверяется гипотеза $H: X_A^T X_A \theta = X_A^T y_A$. Тогда $RSS = (y - X\hat{\theta})^T (y - X\hat{\theta}) = \varepsilon^2$, $RSS_H = \left(y - X\hat{\theta}_A\right)^T \left(y - X\hat{\theta}_A\right)$. Числитель F-статистики будет определяться величиной

$$RSS_H - RSS = \varepsilon^2(A) + \Delta^2(B) - \varepsilon^2 = \left(\hat{\theta} - \hat{\theta}_A\right)^T X^T X \left(\hat{\theta} - \hat{\theta}_A\right),$$

а полностью F-статистика запишется как

$$F = \frac{(\varepsilon^2(A) + \Delta^2(B) - \varepsilon^2)/s}{\varepsilon^2/(N - s)}.$$

Можно рассмотреть также и проверку гипотезы о равенстве регрессий на выборках A и B. Гипотеза записывается как $H:\ \theta_A = \theta_B$, а $F-$ статистика как

$$F = \frac{V^2/s}{(\varepsilon^2(A) + \varepsilon^2(B))/(N - 2s)}.$$

Таким образом, дополнительно к рассмотренным выше в качестве критериев селекции можно использовать функции в виде соответствующих F-статистик проверки гипотез.

Завершая обзор внешних критериев, отметим, что в практике регрессионного моделирования получило распространение использование повторных выборок. В этом случае удается предложить к использованию так называемый субидеальный критерий стабильности [11]. Для многомерных моделей, имеющих векторный отклик, в работах [12–16] предложены и исследованы матричные аналоги критериев селекции, использующие тестовые выборки.

РАЗБИЕНИЕ ВЫБОРКИ ДЛЯ ВНЕШНИХ КРИТЕРИЕВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Использование внешних критериев селекции при решении задачи выбора модели оптимальной сложности предполагает разбиение выборки наблюдения на две части: обучающую и проверочную. На обучающей выборке производится оценивание параметров тестируемых моделей, а на проверочной – проверка их прогнозируемых свойств или свойств согласованности решений с обучающей частью выборки.

В данном разделе основное внимание будет уделено критериям качества моделей, связанным с точностью прогнозирования, в частности – критерию регулярности. В силу этого неизбежно встает задача управления разбиением выборки. Некоторые подходы к решению задачи разбиения с использованием методов оптимального планирования эксперимента предложены в работах [17, 18].

Записывая критерий $\Delta^2(B)$ в канонической форме, легко получить его математическое ожидание [19]:

$$E(\Delta^2(B)) = \left(\widehat{X}_B - P_{BA}\widehat{X}_A\widehat{\theta}\right)^T \left(\widehat{X}_B - P_{BA}\widehat{X}_A\widehat{\theta}\right) +$$

$$+\sigma^2\left(n_B + \operatorname{tr}\left(X_A^T X_A\right)^{-1}\left(X_B^T X_B\right)\right), \tag{4}$$

где

$$P_{BA} = X_B\left(X_A^T X_A\right)^{-1} X_A^T .$$

В работе [7] рассмотрены условия, при которых оптимальная структура, соответствующая минимуму (4), совпадает с истиной структурой $\widehat{s} = m$. Эти условия диктуют «квадратично зависимое» разбиение матрицы X :

$$\rho^2 X_A^T X_A = X_B^T X_B , \tag{5}$$

где ρ^2 – некоторое произвольное число. Точное квадратичное разбиение (5) может иметь место лишь в специально подобранной матрице X, что на практике маловероятно. Кроме того, рекомендации типа (5) не учитывают поведение второго слагаемого в (4). С учетом (5) его можно записать как

$$J_\sigma(s,\ \sigma) = \sigma^2\left(n_B + s/\rho^2\right). \tag{6}$$

Скорость возрастания $J_\sigma(s,\ \sigma)$ в зависимости от σ определяет помехоустойчивость критерия селекции моделей. Ясно, что необходимо выбирать разбиение с возможно большим значением ρ^2 при малой величине n_B.

В общем случае разбиения X на X_A и X_B величина J_σ в соответствии с (4) равна

$$J_\sigma(s,\ \sigma) = \sigma^2\left(n_B + \mathrm{tr}\left(X_A^T X_A\right)^{-1} X_B^T X_B\right). \tag{7}$$

Исследуем возможность минимизации J_σ (7) путем выбора того или иного варианта разбиения X на X_A, X_B при условии, что n_B зафиксировано.

Введем следующие обозначения. Пусть ξ есть непрерывный нормированный план, а $M(\xi)$ – информационная матрица, равная $X_A^T X_A / n_A = \sum_{i=1}^{n} p_i x_i x_i^T$. Далее, пусть X определяет собой множество точек, среди которых необходимо выбрать n_A точек, присвоив им веса́, равные $1/n_A$, а остальным точкам присвоить веса, равные нулю. Оптимальный план ξ^* будем находить как решение следующей экстремальной задачи:

$$\xi^* = Arg\max_{\mathrm{p}} \Psi[M(\xi)]. \tag{8}$$

В качестве функционала $\Psi[M(\xi)]$ будем рассматривать определитель информационной матрицы, что соответствует D-оптимальному планированию эксперимента [20]. Для решения задачи (8) воспользуемся методом проекции градиента функционала $\Psi[M(\xi)]$ на активные гиперплоскости. Пусть вектор $p = (p_1, ..., p_n)^T$ соответствует такому состоянию, что первые его n_A компонентов равны $1/n_A$, а остальные n_B компонентов равны нулю. В этом случае линейное многообразие, на которое проектируется вектор-градиент, будет образовано гиперплоскостями активных ограничений $p_{n_A+1} = 0, ..., p_{n_A+n_B} = 0$, $\sum_{i=1}^{n} p_i = 1$. В соответствии с теоремой Куна–Таккера для того, чтобы точка p^* была решением задачи (8), необходимо и достаточно выполнение условий [18]:

$$\frac{\partial\Psi[M(\xi^*)]}{\partial p_j} = \frac{\partial\Psi[M(\xi^*)]}{\partial p_i}, \quad i, j = n_A + 1, ..., n; \tag{9}$$

$$\sum_{j=1}^{n_A} \frac{\partial \Psi[M(\xi^*)]}{\partial p_j} \frac{1}{n_A} - \frac{\partial \Psi[M(\xi^*)]}{\partial p_{n_A+i}} \geq 0, \ i = 1, ..., \ n_B. \tag{10}$$

Для рассматриваемого функционала $\Psi[M(\xi)]$ компоненты вектора градиента имеют вид

$$\frac{\partial \Psi[M(\xi)]}{\partial p_j} = d(x_j, \xi) = x_j^T M(\xi)^{-1} x_j = \mathrm{tr} M(\xi)^{-1} x_j x_j^T, \tag{11}$$

где $d(x_j, \xi)$ – дисперсия оценки математического ожидания отклика в точке x_j.

По теореме оптимальности [20] в точках D-оптимального плана ξ^* значения $d\left(x_j, \xi^*\right)$ равны s, где s – размерность матрицы $M(\xi)$. С учетом этого (10) можно записать как

$$d\left(x_i, \xi^*\right) \leq \frac{1}{n_A} \sum_{j=1}^{n_A} d\left(x_j, \xi^*\right) = s, \ i = n_A + 1, ..., \ n. \tag{12}$$

Если некоторый план ξ не является D-оптимальным, и для него $p_{n_A+1} = 0, ..., p_{n_A+n_B} = 0$, то

$$d(x_i, \xi) > s, \ i = n_A + 1, ..., \ n. \tag{13}$$

Утверждение 1 [18]. Если на X существует D-оптимальный план ξ^*, такой что $p_1^* = ... = p_{n_A+n_B}^* = 1/n_A, \ p_j^* = 0, \ j = n_A + 1, ..., n$, то для него

$$\mathrm{tr} M^{-1}(\xi^*) X_B^T X_B < \mathrm{tr} M^{-1}(\xi) X_B^T X_B,$$

где ξ – не D-оптимальный план.

Проведенный анализ задачи разбиения X на X_A, X_B позволяет предложить достаточно простую схему действий: для заданного полного плана эксперимента в виде имеющейся выборки решается задача построения D-оптимального плана ξ^* с n_A точками из X. Отметим, что к такому выводу мы приходим и при рассмотрении критерия непротиворечивости [21].

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

$$y_k = m(x_k) + e_k, \ k = 1, \ldots, n, \tag{14}$$

где $e_k \in R$ будем считать независимо и одинаково распределенной ошибкой с $E\left[e_k \mid x = x_k\right] = 0$ и $Var\left[e_k\right] = \sigma^2 < \infty$, $m(x)$ – неизвестная действительная гладкая функция и $E\left[y_k \mid x = x_k\right] = m(x_k)$. Вместо неизвестной функции $m(x)$ будем использовать ее аппроксимацию в виде $f(x) = \omega^T \varphi(x) + b$. Функционал эмпирического риска использования такой аппроксимации

$$R_{\text{emp}}(\omega,\ b) = \frac{1}{n}\sum_{k=1}^{n}\left(\left(\omega^T \varphi(x_k) + b\right) - y_k\right)^2 . \tag{15}$$

Задачу нахождения вектора ω и $b \in R$ можно свести к решению следующей задачи оптимизации [22]:

$$\min_{\omega, b, e} J(\omega,\ e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{k=1}^{n} e_k^2 \tag{16}$$

в предположении, что $y_k = \omega^T \varphi(x_k) + b + e_k$, $k = 1, \ldots, n$. В (16) параметр регуляризации γ отвечает за сложность модели, которая в данном случае определяется нормой вектора ω.

Решение задачи (16) обычно проводят в двойственном пространстве с использованием функционала Лагранжа

$$L(\omega,\ b,\ e,\ \alpha) = J(\omega,\ e) - \sum_{k=1}^{n}\alpha_k \left(\omega^T \varphi(x_k) + b + e_k - y_k\right) \tag{17}$$

с лагранжевыми множителями $\alpha_k \in R$.

Условия оптимальности задаются следующим образом:

$$\begin{cases} \dfrac{dL}{d\omega} = 0 \to \omega = \sum\limits_{k=1}^{n} \alpha_k \varphi(x_k),\ k = 1, \ldots, n; \\[2ex] \dfrac{dL}{db} = 0 \to \sum\limits_{k=1}^{n} \alpha_k = 0,\ k = 1, \ldots, n; \\[2ex] \dfrac{dL}{de_k} = 0 \to \alpha_k = \gamma e_k,\ k = 1, \ldots, n; \\[2ex] \dfrac{dL}{d\alpha_k} = 0 \to \omega^T \varphi(x_k) + b + e_k = y_k,\ k = 1, \ldots, n. \end{cases} \tag{18}$$

После исключения ω и e получаем решение:

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega + \dfrac{1}{\gamma} I_n \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{19}$$

где $y=(y_1,\ldots,y_n)^T$, $1_n=(1,\ldots,1)^T$, $\hat{\alpha}=\left(\hat{\alpha}_1,\ldots,\hat{\alpha}_n\right)$ и $\Omega_{kl}=\varphi(x_k)^T\varphi(x_l)$ для $k,l=1,\ldots,n$. Результирующая LS–SVM модель имеет вид

$$\hat{y}_n(x)=\sum_{k=1}^{n}\hat{\alpha}_k K(x,\,x_k)+\hat{b}, \tag{20}$$

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

$$\hat{b}=\frac{1_n^T\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}y}{1_n^T\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}1_n},\quad \hat{\alpha}=\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}\left(y-1_n\hat{b}\right). \tag{21}$$

В случае выборок большого размера для получения оценок всех параметров вместо обращения матриц в (21) решают систему уравнений (19). Точность получаемого решения (20) во многом определяется настройкой внутренних параметров алгоритма LS–SVM, к числу которых относят параметр регуляризации и параметры ядерных функций. Настройка этих параметров идет, как правило, с использованием внешних критериев качества моделей [22, 23].

РАЗБИЕНИЕ ВЫБОРКИ НА ОБУЧАЮЩУЮ И ТЕСТОВУЮ ЧАСТИ ДЛЯ МЕТОДА LS–SVM

При рассмотрении точности оценивания модели (20) основное внимание будем уделять точности оценивания параметров α, убирая из рассмотрение параметр b через центрирование отклика по схеме $y^*=y-\hat{b}$, как это сделано в (21).

Обозначим оценки параметров α, полученные на обучающей выборке, как

$$\hat{\alpha}_A=\left(\Omega_A+\frac{1}{\gamma}I_{n_A}\right)^{-1}\left(y_B^*\right),$$

где $\Omega_A=K\left(x_i,x_j\right),\ i,j=1,\ldots,n_A$.

Для удобства различения точек обучающей и тестовой выборок будем обозначать координаты точек обучающей выборки через x, а координаты точек тестовой выборки через z. С учетом этого элементы ядерной матрицы Φ_B для вычисления прогноза в точке тестовой выборки будем обозначать как

$$(\Phi_B)_{ij}=K(z_i,x_j),\ i=1,\ldots,n_B,\ j=1,\ldots,n_A.$$

Прогнозные значения по модели, полученной на выборке A, рассчитываются так:

$$\hat{y}_B = \Phi_B \hat{\alpha}_A + \hat{b}_A .$$

Ковариационная матрица ошибок прогноза на выборку B имеет вид

$$\operatorname{cov}(\hat{y}_B) = \left(\sigma^2 + \operatorname{cov}\left(\hat{b}_A\right)\right)\Phi_B \left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-2} \Phi_B^T + \operatorname{cov}\left(\hat{b}_A\right),$$

где

$$\operatorname{cov}\left(\hat{b}_A\right) = \sigma^2 \frac{1_{n_A}^T \left(\Omega + \frac{1}{\gamma} I_{n_A}\right)^{-2} 1_{n_A}}{\left[1_{n_A}^T \left(\Omega + \frac{1}{\gamma} I_{n_A}\right)^{-1} 1_{n_A}\right]^2} .$$

Средняя дисперсия прогноза вычисляется так:

$$\overline{\sigma}^2 (\hat{y}_B) = \left(\sigma^2 + \operatorname{cov}\left(\hat{b}_A\right)\right) tr \left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-2} \Phi_B^T \Phi_B / n_B + \operatorname{cov}\left(\hat{b}_A\right).$$

Опираясь на утверждение 1, минимизировать среднюю дисперсию $\overline{\sigma}^2 (\hat{y}_B)$ будем опосредовано через минимизацию определителя дисперсионной матрицы оценок параметров α. В нашем случае эта дисперсионная матрица имеет вид

$$\operatorname{cov}(\hat{\alpha}_A) = \left(\sigma^2 + \operatorname{cov}\left(\hat{b}_A\right)\right)\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-2} . \qquad (22)$$

Поведение $\operatorname{cov}\left(\hat{b}_A\right)$ можно рассмотреть на следующем примере. В качестве ядерной функции возьмем гауссову (RBF-ядро). В матрице Ω все диагональные элементы $k(x_i, x_i) = 1$. Недиагональные элементы неотрицательны. Сумма всех элементов матрицы $\left(\Omega + \frac{1}{\gamma} I_{n_A}\right)^{-1}$ достигает минимального значения в том числе, когда ее недиагональные элементы близки к нулю. Это возможно, когда параметр масштаба гауссовой ядерной функции выбран достаточно большим или когда точки выборки расположены на достаточно большом расстоянии друг от друга. Будем считать, что это так, тогда $\operatorname{cov}\left(\hat{b}_A\right) \cong \sigma^2 \frac{n_A(\gamma / (1+\gamma))^2}{n_A^2(\gamma / (1+\gamma))^2} = \frac{\sigma^2}{n_A}$. Именно такой дисперсией обладает оценка параметра b в виде среднего $\tilde{b} = 1^T y / n_A$. Учитывая, что

$\left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-2}\right| = \left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1}\right| * \left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1}\right|$ и то, что матрица $\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1}$ положительно определена, будем рассматривать минимизацию определителя $\left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1}\right|$ или, что намного проще, – максимизацию определителя $\left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)\right|$. Для определителя положительно определенной матрицы известно свойство, что он меньше либо равен произведению диагональных ее элементов. Поскольку все диагональные элементы матрицы $\Omega_A + \frac{1}{\gamma} I_{n_A}$ равны $(\gamma + 1)/\gamma$, можно заключить, что максимум определителя достигается при диагональной матрице Ω_A. При этом $\mathrm{cov}\left(\hat{b}_A\right) = \frac{\sigma^2}{n_A}$. Таким образом, в целях упрощения задачи минимизации определителя матрицы ковариации (22) будем решать задачу максимизации определителя $\left|\left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)\right|$. Тем самым мы будем строить дискретный *D*-оптимальный план объемом в n_A наблюдений, используя все точки имеющейся выборки.

В нашем случае для построения дискретного *D*-оптимального плана удобно воспользоваться хорошо себя зарекомендовавшими последовательными алгоритмами [24, 25].

Обозначим через G_s матрицу размером $s \times s$ для обучающей выборки объемом в s наблюдений и состоящую из элементов $(G_s)_{ij} = K(x_i,\, x_j) + \frac{1}{\gamma} I_s,\ i, j = 1, \ldots, s$.

Тогда на шаге $s+1$ матрица G_{s+1} будет иметь вид

$$G_{s+1} = \begin{pmatrix} G_s & F(x_{s+1}) \\ F^T(x_{s+1}) & K(x_{s+1}, x_{s+1}) + \frac{1}{\gamma} \end{pmatrix},$$

где $F^T(x_{s+1}) = (K(x_1, x_{s+1}),\ K(x_2, x_{s+1}), \ldots,\ K(x_s, x_{s+1}))$.

Определитель окаймленной матрицы легко вычисляется:

$$|G_{s+1}| = |G_s| * \Delta(x_{s+1}),$$

где $\Delta(x_{s+1}) = \left[K(x_{s+1},\, x_{s+1}) + \frac{1}{\gamma} - F^T(x_{s+1}) G_s^{-1} F(x_{s+1})\right]$.

Таким образом, очередная точка, включаемая в обучающую выборку, отыскивается по следующей схеме: $x_{s+1} = \operatorname{Arg}\max_{x} \Delta(x)$, где аргумент x принимает значения координат точек исходной выборки, еще не включенных в обучающую часть.

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Разбиение выборки на обучающую и тестовую части, как мы уже отмечали, можно использовать для настройки параметров ядерных функций с целью получения решения с хорошей обобщающей способностью. Целью вычислительного эксперимента являлось сравнение полученных решений с разбиением выборки на тестовую и обучающую части случайным образом и с помощью D-оптимального планирования. В качестве критериев точности моделей использовался критерий регулярности и критерий скользящего прогноза (CV).

Для проведения исследования использовалась тестовая функция: $m(x) = 7 / e^{(x+0.75)^2} + 3x$, заданная на отрезке [1; 1]. В качестве ядерной функции использовалось RBF-ядро. В качестве помехи использовались нормально распределенные величины. Уровень помехи (дисперсия случайной величины) выбирался как 5 % и 20 % от мощности незашумленного сигнала. Количество наблюдений выбиралось равным 10, 20 и 30. При проведении вычислительных экспериментов параметр регуляризации γ принимал фиксированное значение 10. Подбор лучшего решения осуществлялся по параметру масштаба RBF-ядра, который варьировался от 10^{-1} до 10^{1} с шагом 0,1.

В табл. 1–3 приведены усредненные по 30 реализациям шума значения среднеквадратичной ошибки (MSE) для метода LS–SVM с RBF-ядром. В таблицах в столбцах, озаглавленных как CV, $\Delta^2_{\text{reg}}(S)$, $\Delta^2_{\text{reg}}(D)$, представлены соответственно средние значения MSE, полученные при использовании критерия скользящего контроля, критерия регулярности на случайной тестовой выборке и критерия регулярности на тестовой выборке при D-опти-мальном разбиении. Условия экспериментов по строкам различались тем, что использовались различные варианты случайного разбиения выборки на обучающую и тестовую части. При этом реализации шума также различались. Разбиение выборки на обучающую и тестовую части с использованием D-оптимального планирования эксперимента фиксировалось одним и тем же. Различие результатов по какому-либо столбцу $\Delta^2_{\text{reg}}(D)$ обусловлено использованием различных наборов реализаций шума. Это замечание справедливо и к результатам, отраженным в столбце CV.

Таблица 1

Средние значения критерия MSE для выборки объемом 10 наблюдений при уровне шума 20 %

Номер варианта разбиения	CV	Количество наблюдений в тестовой части в %							
		10		15		20		25	
		$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$
1	0,024	0,014	0,041	0,019	0,019	0,019	0,019	0,013	0,019
2	0,028	0,165	0,032	0,020	0,016	0,020	0,016	0,019	0,016
3	0,024	0,012	0,026	0,021	0,018	0,021	0,018	0,013	0,017
4	0,027	0,015	0,040	0,015	0,023	0,015	0,023	0,016	0,025
5	0,025	0,039	0,014	0,022	0,024	0,022	0,024	0,017	0,021
6	0,023	0,013	0,045	0,021	0,033	0,021	0,033	0,115	0,021
7	0,022	0,051	0,023	0,014	0,014	0,014	0,014	0,014	0,015
8	0,024	0,028	0,028	0,018	0,028	0,018	0,028	0,015	0,026
9	0,022	0,164	0,031	0,035	0,024	0,035	0,024	0,013	0,022
10	0,022	0,088	0,034	0,060	0,030	0,060	0,030	0,016	0,021
11	0,024	0,137	0,036	0,038	0,022	0,038	0,022	0,033	0,021
12	0,025	0,059	0,031	0,018	0,027	0,018	0,027	0,014	0,028
13	0,024	0,168	0,013	0,028	0,019	0,028	0,019	0,013	0,020
14	0,023	0,141	0,045	0,030	0,025	0,030	0,025	0,014	0,017
15	0,026	0,118	0,034	0,168	0,022	0,168	0,022	0,030	0,020
16	0,024	0,206	0,040	0,013	0,020	0,013	0,020	0,023	0,020
17	0,024	0,182	0,034	0,055	0,018	0,055	0,018	0,024	0,017
18	0,024	0,019	0,037	0,045	0,019	0,045	0,019	0,014	0,019
19	0,025	0,013	0,047	0,014	0,018	0,014	0,018	0,038	0,020
20	0,022	0,103	0,049	0,019	0,029	0,019	0,029	0,025	0,021

Таблица 2

Средние значения MSE для выборки объемом 20 наблюдений при уровне шума 20 %

Номер варианта разбиения	CV	Количество наблюдений в тестовой части в %							
		10		15		20		25	
		$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$
1	0,009	0,017	0,010	0,011	0,011	0,011	0,008	0,015	0,009
2	0,009	0,037	0,015	0,007	0,010	0,007	0,008	0,011	0,008
3	0,009	0,016	0,010	0,009	0,011	0,008	0,009	0,009	0,008
4	0,008	0,013	0,009	0,006	0,009	0,014	0,006	0,012	0,006
5	0,009	0,007	0,014	0,009	0,010	0,009	0,009	0,007	0,008
6	0,009	0,010	0,013	0,006	0,011	0,006	0,008	0,012	0,008
7	0,009	0,009	0,011	0,007	0,009	0,007	0,008	0,009	0,008
8	0,010	0,007	0,020	0,011	0,011	0,007	0,009	0,011	0,010
9	0,008	0,023	0,015	0,006	0,010	0,012	0,008	0,009	0,009
10	0,008	0,012	0,012	0,009	0,009	0,016	0,008	0,009	0,008
11	0,009	0,033	0,013	0,018	0,013	0,010	0,009	0,007	0,009
12	0,008	0,014	0,010	0,016	0,009	0,006	0,008	0,011	0,008
13	0,010	0,015	0,014	0,013	0,011	0,010	0,009	0,007	0,007
14	0,009	0,014	0,012	0,008	0,010	0,007	0,009	0,009	0,007
15	0,010	0,011	0,013	0,008	0,009	0,014	0,009	0,009	0,009
16	0,009	0,007	0,010	0,008	0,009	0,007	0,009	0,009	0,007
17	0,008	0,018	0,010	0,011	0,010	0,013	0,008	0,007	0,008
18	0,009	0,085	0,020	0,008	0,011	0,011	0,008	0,007	0,008
19	0,009	0,012	0,009	0,037	0,010	0,011	0,008	0,007	0,009
20	0,009	0,013	0,010	0,014	0,010	0,008	0,008	0,011	0,008

Таблица 3

Средние значения MSE для выборки объемом 30 наблюдений при уровне шума 20 %

Номер варианта разбиения	CV	Количество наблюдений в тестовой части в %							
		10		15		20		25	
		$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$	$\Delta^2_{\text{reg}}(S)$	$\Delta^2_{\text{reg}}(D)$
1	0,005	0,011	0,008	0,004	0,007	0,004	0,005	0,006	0,006
2	0,006	0,012	0,008	0,007	0,008	0,005	0,006	0,009	0,006
3	0,006	0,006	0,008	0,005	0,006	0,007	0,007	0,005	0,006
4	0,006	0,014	0,006	0,006	0,007	0,008	0,007	0,007	0,005
5	0,005	0,005	0,006	0,005	0,006	0,005	0,005	0,007	0,006
6	0,005	0,011	0,007	0,006	0,006	0,005	0,006	0,007	0,006
7	0,005	0,005	0,007	0,005	0,006	0,005	0,007	0,006	0,006
8	0,006	0,007	0,008	0,008	0,008	0,007	0,007	0,005	0,006
9	0,005	0,005	0,008	0,006	0,007	0,005	0,007	0,004	0,006
10	0,006	0,007	0,007	0,008	0,006	0,008	0,007	0,005	0,007
11	0,006	0,006	0,005	0,007	0,006	0,005	0,006	0,007	0,006
12	0,006	0,008	0,007	0,009	0,007	0,005	0,006	0,005	0,006
13	0,005	0,004	0,007	0,007	0,006	0,006	0,006	0,006	0,006
14	0,005	0,011	0,009	0,005	0,006	0,004	0,005	0,006	0,006
15	0,005	0,010	0,007	0,005	0,007	0,006	0,005	0,006	0,006
16	0,005	0,008	0,007	0,007	0,006	0,005	0,005	0,006	0,006
17	0,006	0,005	0,007	0,006	0,007	0,005	0,006	0,005	0,006
18	0,005	0,005	0,006	0,005	0,006	0,004	0,006	0,004	0,005
19	0,006	0,007	0,009	0,006	0,007	0,010	0,007	0,005	0,007
20	0,005	0,006	0,007	0,005	0,006	0,008	0,006	0,005	0,006

По представленным в табл. 1–3 средним значениям MSE можно судить об устойчивости получаемых решений при использовании того или иного критерия качества. Как и следовало ожидать, качество решений при использовании критерия регулярности при случайном разбиении выборки на обучающую и тестовую части во многом зависит от самого варианта разбиения. Это наглядно представлено на рис. 1–2, на которых ось абсцисс – номер варианта случайного разбиения.

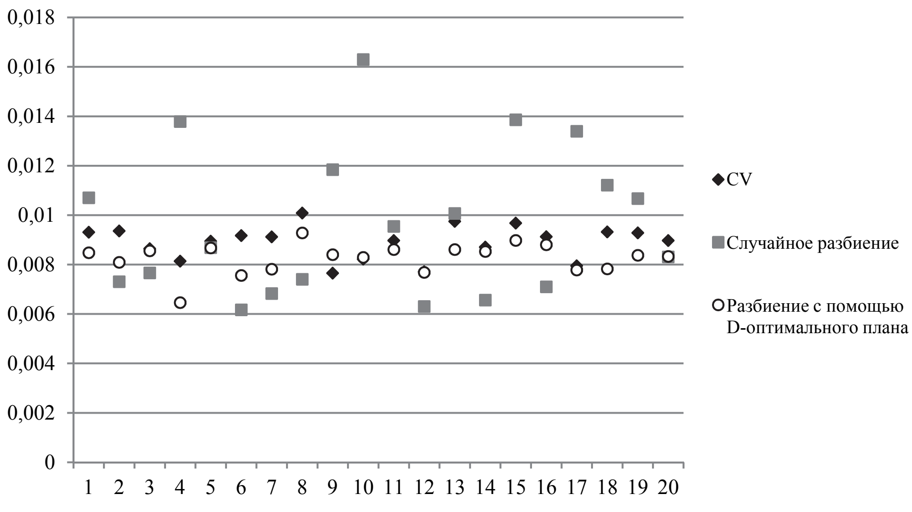

Рис. 1. График средних значений MSE по 30 реализациям для выборки объемом 20 с количеством наблюдений 20 % в тестовой части (уровень шума 20 %)

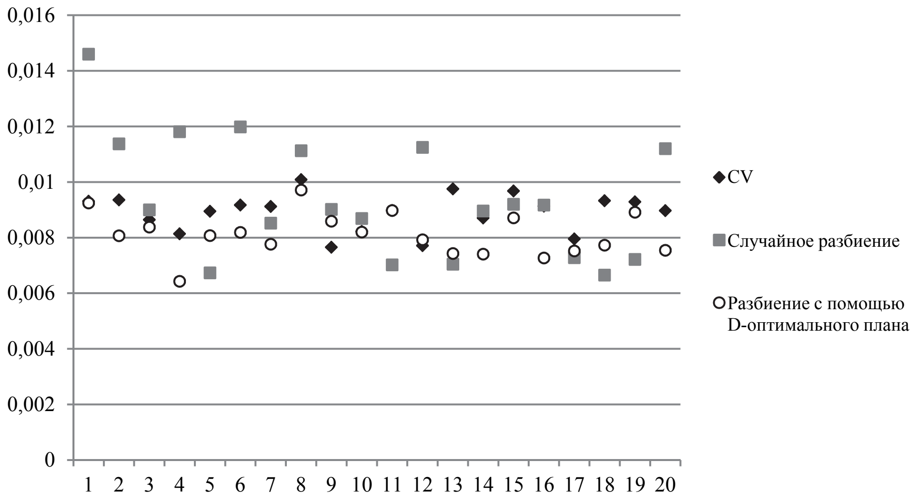

Рис. 2. График средних значений MSE по 30 реализациям для выборки объемом 20 с количеством наблюдений 25 % в тестовой части (уровень шума 20 %)

ЗАКЛЮЧЕНИЕ

В работе для получения модели оптимальной сложности на основе LS–SVM предложен способ разбиения выборки на тестовую и обучающую части с использованием метода планирования эксперимента. Для получения *D*-оптимального разбиения предложен последовательный алгоритм.

По результатам проведенных вычислительных экспериментов можно сделать выводы о том, что эффективность использования случайной тестовой выборки нестабильна и во много определяется конкретным вариантом разбиения. При этом стабильность результатов использования тестовой выборки, полученной при *D*-оптимальном разбиении выборки, значительно выше. Качество получаемых решений при использовании для настройки параметров критериев CV и регулярности на *D*-оптимальной тестовой выборке близкó. Таким образом, для решения задачи настройки параметров алгоритма LS–SVM можно рекомендовать использовать разбиение выборки на обучающую и тестовую части по методу *D*-оптимального планирования эксперимента.

СПИСОК ЛИТЕРАТУРЫ

1. *Перельман И.И.* Методология выбора структуры модели при идентификации объектов управления // Автоматика и телемеханика. – 1983. – № 11. – С. 5–29.

2. *Романов В.Л.* Выбор наилучшей линейной регрессии: сравнение формальных критериев // Заводская лаборатория. – 1990. – № 1. – С. 90–95.

3. *Себер Дж.* Линейный регрессионный анализ. – М.: Мир, 1980. – 456 с.

4. *Степашко В.С., Кочерга Ю.Л.* Методы и критерии решения задач структурной идентификации // Автоматика. – 1985. – № 5. – С. 29–37.

5. *Кочерга Ю.Л.* J-оптимальная редукция структуры модели в схеме Гаусса–Маркова // Автоматика. – 1988. – № 4. – С. 34–38.

6. *Сарычев А.П.* Усредненный критерий регулярности метода группового учета аргументов в задаче поиска наилучшей регрессии // Автоматика. – 1990. – № 5. – С. 28–33.

7. *Степашко В.С.* Асимптотические свойства внешних критериев выбора моделей // Автоматика. – 1988. – № 6. – С. 75–82.

8. *Степашко В.С.* Потенциальная помехоустойчивость моделирования по комбинаторному алгоритму МГУА без использования информации о помехах // Автоматика. – 1983. – № 3. – С. 18–28.

9. *Степашко В.С.* Селективные свойства критерия непротиворечивости моделей // Автоматика. – 1986. – № 2. – С. 40–49.

10. *Попов А.А.* Методы планирования эксперимента в задачах синтеза моделей оптимальной сложности // Машинные методы планирования эксперимента и оптимизации многофакторных систем / Новосибирский электротехнический институт. – Новосибирск, 1987. – С. 54–58.

11. *Попов А.А.* Использование повторных выборок в критериях селекции моделей // Планирование эксперимента, идентификация, анализ и оптимизация многофакторных систем / Новосибирский электротехнический институт. – Новосибирск, 1990. – С. 82–88.

12. *Лисицин Д.В., Попов А.А.* Выбор структуры для многомерной динамической системы // Сборник научных трудов НГТУ. – 1997. – Вып. 1 (6). – С. 33–40.

13. *Лисицин Д.В., Попов А.А.* Исследование критериев селекции многомерных моделей при наличии разнотипных факторов // Труды третьей международной научно-технической конференции «Актуальные проблемы электронного приборостроения» АПЭП-96. – Новосибирск, 1996. – Т. 6, ч. 1. – С. 54–58.

14. *Лисицин Д.В., Попов А.А.* Исследование критериев селекции многооткликовых регрессионных моделей // Сборник научных трудов НГТУ. – 1996. – Вып. 2. – С. 19–28.

15. *Лисицин Д.В., Попов А.А.* Конструирование критериев селекции многомерных регрессионных моделей // Сборник научных трудов НГТУ. – 1996. – Вып. 1. – С. 13–20.

16. *Лисицин Д.В., Попов А.А.* Структурная оптимизация многомерных регрессионных моделей // Второй Сибирский Конгресс по прикладной и индустриальной математике: тезисы докладов. – Новосибирск, 1996. – С. 179.

17. *Попов А.А.* Планирование эксперимента в задачах разбиения выборки в МГУА // Сборник научных трудов НГТУ. – 1995. – Вып. 2. – С. 35–40.

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

19. *Юрачковский Ю.П., Грошков А.Н.* Применение канонической формы внешних критериев для исследования их свойств // Автоматика. – 1979. – № 3. – С. 85–89.

20. *Федоров В.В.* Активные регрессионные эксперименты // Математические методы планирования эксперимента. – Новосибирск: Наука, 1981. – С. 19–73.

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

22. Least square support vector machines / J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle. – New Jersey: World Scientific, 2002. – 290 p.

23. *Попов А.А., Бобоев Ш.А.* Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов // Сборник научных трудов НГТУ. – 2015. – № 3 (81). – С. 69–78.

24. *Попов А.А.* Последовательные схемы построения оптимальных планов эксперимента // Сборник научных трудов НГТУ. – 1995. – Вып. 1. – С. 39–44.

25. *Попов А.А.* Последовательные схемы синтеза оптимальных планов эксперимента // Доклады АН ВШ РФ. – 2008. – № 1 (10). – С. 45–55.

Попов Александр Александрович, доктор технических наук, профессор кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основные направления научных исследований: статистические методы анализа данных, оптимальное планирование экспериментов. Имеет более 150 публикаций, в том числе 3 монографии. E-mail: a.popov@corp.nstu.ru

Бобоев Шараф Асрорович, аспирант кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление исследований – статистические методы анализа данных. Имеет 4 публикаций. E-mail: shboboev@mail.ru

Obtaining a test sample by the LS–SVM method using optimal experiment planning[*]

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, a post-graduate student. E-mail: shboboev@mail.ru*

The problem of regression dependence recovery by the method of support vector machines with a quadratic loss function is studied in the paper. This method belongs to the kernel technique class. To set a number of LS–SVM algorithm internal parameters the problem of obtaining a test sample is discussed. Various criterions of model selection which are based on partitioning the sample into learning and test parts are presented. The problem of partitioning the sample into learning and test parts with the use of the *D*-optimal experiment planning method is considered in detail for the case of linear parametric regression models. This method of obtaining test samples is proposed to use for the LS–SVM method. A sequential algorithm is presented for obtaining the learning and test parts of the sample observations as applied to the LS–SVM method. To verify the efficiency of the proposed method of partitioning samples a computational experiment was conducted. An improvement of the LS–SVM solution accuracy was achieved by selecting the scale of the Gaussian kernel function. This parameter of the kernel function was selected by minimizing a prediction error in the sample test part. The final solution accuracy was tested by the mean-square error method. The computational experiment was carried out on simulated data. A nonlinear dependence on the input factor was selected as a data generating model. The variance of the noise (noise level) was determined as a percentage of a signal power. Two ways of sample partitioning were compared on the learning and test parts, namely random partitioning and partitioning by the *D*-optimal experiment planning method. The cross-validation criterion was also used to select the LS–SVM algorithm parameters. The results of computational experiments are given in tables and figures. Based on the results of the computational experiments, conclusions are made that the results of using a random test sample are unstable and largely depend on the specific partitioning option. Whereas the stability of test sample results obtained by *D*-optimal sample partitioning is much higher.

Keywords: regression, LS-SVM method, quadratic loss function, test sample training sample, optimal experiment planning, *D*-optimal plan, regularity criterion, cross-validation criterion, regularization coefficient, kernel function, mean square error

DOI: 10.17212/1814-1196-2016-4-80-99

REFERENCES

1. Perel'man I.I. Metodologiya vybora struktury modeli pri identifikatsii ob"ektov upravleniya [A methodology for the selection of the model structure when identification of objects of management]. *Avtomatika i telemekhanika – Automation and Remote Control*, 1983, no. 11, pp. 5–29. (In Russian)

2. Romanov V.L. Vybor nailuchshei lineinoi regressii: sravnenie formal'nykh kriteriev [The selection of the best linear regression: a comparison of formal criteria]. *Zavodskaya laboratoriya – Industrial Laboratory,* 1990, no. 1, pp. 90–95. (In Russian)

3. Seber G.A.F. *Linear regression analysis*. New York, Wiley, 1977 (Russ. ed.: Seber Dzh. *Lineinyi regressionnyi analiz*. Moscow, Mir Publ., 1980. 456 p.).

4. Stepashko V.S., Kocherga Yu.L. Metody i kriterii resheniya zadach strukturnoi identifikatsii [Methods and criteria of the solving problems of structural identification]. *Avtomatika – Soviet Journal of Automation and Information Sciences*, 1985, no. 5, pp. 29–37. (In Russian)

[*] *Received 30 August 2016.*

5. Kocherga Yu.L. J-optimal'naya reduktsiya struktury modeli v skheme Gaussa–Markova [J-optimal reduction of structure of model in the scheme of Gauss–Markov]. *Avtomatika – Soviet Journal of Automation and Information Sciences*, 1988, no. 4, pp. 34–38. (In Russian)

6. Sarychev A.P. Usrednennyi kriterii regulyarnosti metoda gruppovogo ucheta argumentov v zadache poiska nailuchshei regressii [The averaged regularity criterion of group method of accounting arguments in the problem of finding the best regression]. *Avtomatika – Soviet Journal of Automation and Information Sciences*, 1990, no. 5, pp. 28–33. (In Russian)

7. Stepashko V.S. Asimptoticheskie svoistva vneshnikh kriteriev vybora modelei [The asymptotic properties of the external criteria of selection models]. *Avtomatika – Soviet Journal of Automation and Information Sciences*, 1988, no. 6, pp. 75–82. (In Russian)

8. Stepashko V.S. Potentsial'naya pomekhoustoichivost' modelirovaniya po kombinatornomu algoritmu MGUA bez ispol'zovaniya informatsii o pomekhakh [The potential noise immunity of modeling by combinatorial GMDH algorithm without using the interference information]. *Avtomatika – Soviet Automatic Control*, 1983, no. 3, pp. 18–28. (In Russian)

9. Stepashko V.S. Selektivnye svoistva kriteriya neprotivorechivosti modelei [The selective properties of the consistency criterion of models]. *Avtomatika – Soviet Journal of Automation and Information Sciences*, 1986, no. 2, pp. 40–49. (In Russian)

10. Popov A.A. [The methods of experimental planning in problems of optimal complexity models synthesis]. *Mashinnye metody planirovaniya eksperimenta i optimizatsii mnogofaktornykh sistem* [The machine methods of experimental planning and optimization of multifactor systems]. Novosibirsk electrotechnical institute. Novosibirsk, 1987, pp. 54–58.

11. Popov A.A. [The use of repeated samples in the criteria of selection models]. *Planirovanie eksperimenta, identifikatsiya, analiz i optimizatsiya mnogofaktornykh sistem* [Experiment planning, identification, analysis and optimization of multifactor systems]. Novosibirsk electrotechnical institute. Novosibirsk, 1990, pp. 82–88.

12. Lisitsin D.V., Popov A.A. Vybor struktury dlya mnogomernoi dinamicheskoi sistemy [The selection of structure for multidimensional dynamic systems]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1997, no. 1 (6), pp. 33–40.

13. Lisitsin D.V., Popov A.A. [A researching of selection criteria of multidimensional models in the presence of different types of factors]. *Trudy III mezhdunarodnoi nauchno-tekhnicheskoi konferentsii "Aktual'nye problemy elektronnogo priborostroeniya" APEP-96* [Proceedings of Third international scientific-technical conference "Actual problems of electronic instrument engineering" APEIE-96]. Novosibirsk, 1996, vol. 6, pt. 1, pp. 54–58.

14. Lisitsin D.V., Popov A.A. Issledovanie kriteriev selektsii mnogootklikovykh regressionnykh modelei [A researching of selection criteria of multiresponse regression models]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1996, no. 2, pp. 19–28.

15. Lisitsin D.V., Popov A.A. Konstruirovanie kriteriev selektsii mnogomernykh regressionnykh modelei [The development of selection criteria of multidimentional 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.

16. Lisitsin D.V., Popov A.A. [The structural optimization of multidimentional regression models]. *Vtoroi Sibirskii Kongress po prikladnoi i industrial'noi matematike: tezisy dokladov* [The Second Siberian Congress on Industrial and Applied Mathematics: abstracts]. Novosibirsk, 1996, p. 179.

17. Popov A.A. Planirovanie eksperimenta v zadachakh razbieniya vyborki v MGUA [The experiment planning in problems of splitting the sample in GMDH]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1995, no. 2, pp. 35–40.

18. Popov A.A. Razbienie vyborki dlya vneshnikh kriteriev selektsii modelei s ispol'zovaniem metodov planirovaniya eksperimenta [The splitting of sample for the external criteria of selection models using experiment planning methods]. *Zavodskaya laboratoriya. Diagnostika materialov – Industrial laboratory. Materials diagnostics*, 1997, no. 1, pp. 49–53. (In Russian)

19. Yurachkovskii Yu.P., Groshkov A.N. Primenenie kanonicheskoi formy vneshnikh kriteriev dlya is-sledovaniya ikh svoistv [The use of canonical form of external criteria for the research of their properties]. *Avtomatika – Soviet Automatic Control*, 1979, no. 3, pp. 85–89. (In Russian)

20. Fedorov V.V. [The active regression experiments]. *Matematicheskie metody planirovaniya eksperimenta* [The mathematical methods of experimental planning]. Novosibirsk, Nauka Publ., 1987, pp. 19–73.

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

22. Suykens J.A.K., Gestel T. van, Brabanter J. de, Moor B. de, Vandewalle J. *Least square support vector machines*. New Jersey, World Scientific, 2002. 290 p.

23. Popov A.A., Boboev Sh.A. Postroenie regressionnykh zavisimostei s ispol'zovaniem kvadratichnoi funktsii poter' v metode opornykh vektorov [The construction of a regression relationships using least square in support vector machines]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 2015, no. 3 (81), pp. 69–78.

24. Popov A.A. Posledovatel'nye skhemy postroeniya optimal'nykh planov eksperimenta [The sequential schemes constructing of the optimal experiment plans]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1995, no. 1, pp. 39–44.

25. Popov A.A. Posledovatel'nye skhemy sinteza optimal'nykh planov eksperimenta [Sequential schemes of synthesis of optimum plans of experiment]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2008, no. 1 (10), pp. 45–55.