КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2016 Т. 8 № 3 С. 475–484

Ки&М

ЧИСЛЕННЫЕ МЕТОДЫ И ОСНОВЫ ИХ РЕАЛИЗАЦИИ

УДК: 519.6

Сокращение вида решающего правила метода многомерной интерполяции и аппроксимации в задаче классификации данных

И. В. Копылов

ООО «Малленом системс»,
Россия, 162610, г. Череповец, ул. Металлургов, д. 21б

E-mail: ivv.kopylov@gmail.com

Получено 18.03.2016.
Принято к публикации 28.04.2016.

В данной статье исследуется метод машинного обучения на основе теории случайных функций. Одной из основных проблем данного метода является то, что вид решающего правила модели метода, построенной на данных обучающей выборки, становится более громоздким при увеличении количества примеров выборки. Решающее правило модели является наиболее вероятной реализацией случайной функции и представляется в виде многочлена с количеством слагаемых, равным количеству обучающих элементов выборки. В статье будет показано, что для рассматриваемого метода существует быстрый способ сокращения обучающей выборки и, соответственно, вида решающего правила. Уменьшение примеров обучающей выборки происходит за счет поиска и удаления малоинформативных (слабых) элементов, которые незначительно влияют на итоговый вид решающей функции, и шумовых элементов выборки. Для каждого (x_i, y_i)-го элемента выборки было введено понятие значимости, выражающееся величиной отклонения оцененного значения решающей функции модели в точке x_i, построенной без *i*-го элемента, от реального значения y_i. Будет показана возможность косвенного использования найденных слабых элементов выборки при обучении модели метода, что позволяет не увеличивать количество слагаемых в полученной решающей функции. Также в статье будут описаны проведенные эксперименты, в которых показано, как изменение количества обучающих данных влияет на обобщающую способность решающего правила модели в задаче классификации.

Ключевые слова: машинное обучение, интерполяция, аппроксимация, случайная функция, система линейных уравнений, скользящий контроль, классификация

COMPUTER RESEARCH
AND MODELING 2016 VOL. 8 NO. 3 P. 475–484

Ки&М

NUMERICAL METHODS AND THE BASIS FOR THEIR APPLICATION

UDC: 519.6

Reduction of decision rule of multivariate interpolation and approximation method in the problem of data classification

I. V. Kopylov

Ltd. «Mallenom Systems»
21b Metallurgov st., Cherepovets, 162610, Russia

E-mail: ivv.kopylov@gmail.com

Retrieved 18.03.2016.
Accepted for publication 28.04.2016.

This article explores a method of machine learning based on the theory of random functions. One of the main problems of this method is that decision rule of a model becomes more complicated as the number of training dataset examples increases. The decision rule of the model is the most probable realization of a random function and it's represented as a polynomial with the number of terms equal to the number of training examples. In this article we will show the quick way of the number of training dataset examples reduction and, accordingly, the complexity of the decision rule. Reducing the number of examples of training dataset is due to the search and removal of weak elements that have little effect on the final form of the decision function, and noise sampling elements. For each (x_i, y_i)-th element sample was introduced the concept of value, which is expressed by the deviation of the estimated value of the decision function of the model at the point x_i, built without the i-th element, from the true value y_i. Also we show the possibility of indirect using weak elements in the process of training model without increasing the number of terms in the decision function. At the experimental part of the article, we show how changed amount of data affects to the ability of the method of generalizing in the classification task.

Keywords: machine learning, interpolation, approximation, random function, the system of linear equations, cross-validation, classification

Citation: *Computer Research and Modeling*, 2016, vol. 8, no. 3, pp. 475–484 (Russian).

Введение

Рассмотрим основные моменты, связанные с методом машинного обучения на основе теории случайных функций [Бахвалов, Зуев, Ширабакина, 2005; Бахвалов, Малыгин, Черкас, 2012]. Все формулы будут адаптированы для задачи классификации.

Пусть последовательность $x_1, x_2, \cdots, x_k \left(x_i \in R^n\right)$ и соответствующие им $y_1, y_2, \cdots, y_k \left(y_i \in R^m\right)$ представляют собой обучающую выборку: $x_i = \left(x_{i1}, x_{i2}, \cdots, x_{in}\right)$ — вектор признаков, описывающих *i*-й элемент обучающей выборки; $y_i = \left(y_{i1}, y_{i2}, \cdots, y_{im}\right)$ — выходной вектор. В случае задачи классификации для x_i, примера обучающей выборки, y_i будет результирующим вектором, хранящим информацию, к какому классу относится x_i. Все элементы вектора y_i равны нулю, кроме одного, который равен 1 (соответствует классу *i*-го примера).

Решающие правила метода в общем виде можно представить следующим образом:

$$f_j(x) = q_{1j} K_f(x - x_1) + q_{2j} K_f(x - x_2) + \cdots + q_{kj} K_f(x - x_k), \tag{1}$$

где $q_{i,j} \in R$, причем $arg \max_j \left(f_j(x)\right)$ будет предсказанным классом для x.

$K_f(\tau)$ задается формулой (2):

$$K_f(\tau) = C_K \left((\tau_2)^2 \ln\left(\frac{(\tau_2)^2}{t} \right) + d \right), \tag{2}$$

где C_K, t, d — расчетные коэффициенты, выведенные в рамках теории рассматриваемого метода; τ_2 — норма вектора, $\tau \in R^n$. Коэффициенты $t \approx 10^5$ и $d \approx 10^5$ (при условии нормирования входных значений в диапазоне [–1, 1]). Коэффициент C_K является калибровочным, связан со свойствами априорной случайной функции (при неизвестном уровне дисперсии значений возможных реализаций случайной функции на единичном расстоянии от какого-либо элемента обучения C_K принимает значение 1). Функция K_f является симметричной и характеризует зависимость ожидаемого различия между значениями функции f в некоторых двух точках от расположения этих точек.

Вместо формулы (2) может быть иная симметричная функция и могут рассматриваться другие методы машинного обучения, например метод на основе обратно взвешенных расстояний, метод радиальных базисных функций, кригинг [Robeson, 1997; Кошель, Мусин, 2000].

Коэффициенты q_{ij} $(i = 1, \cdots, k; j = 1, \cdots, m)$ находятся решением системы линейных уравнений (3):

$$\begin{cases} q_{1j}\left(K_f(x_1 - x_1) + S(x_1)\right) + q_{2j} K_f(x_1 - x_2) + \cdots + q_{kj} K_f(x_1 - x_k) = y_{1j}, \\ q_{1j} K_f(x_2 - x_1) + q_{2j}\left(K_f(x_2 - x_2) + S(x_2)\right) + \cdots + q_{kj} K_f(x_2 - x_k) = y_{2j}, \\ \vdots \\ q_{1j} K_f(x_k - x_1) + q_{2j} K_f(x_k - x_2) + \cdots + q_{kj}\left(K_f(x_k - x_k) + S(x_k)\right) = y_{kj}. \end{cases} \tag{3}$$

Значения $S(x)$ определяют априорно предполагаемый уровень шума (погрешности) в данных обучения и, соответственно, степень приближения, с которой функция (1) воспроизведет данные обучения.

Перепишем (3) в матричном виде:

$$KQ = Y, \tag{4}$$

где Q и Y — вектор-столбцы, а K — симметричная матрица:

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,m} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ q_{k,1} & q_{k,2} & \cdots & q_{k,m} \end{bmatrix}_{k\times m}, \tag{5}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,m} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ y_{k,1} & y_{k,2} & \cdots & y_{k,m} \end{bmatrix}_{k\times m}, \tag{6}$$

$$K = \begin{bmatrix} K_f(x_1 - x_1) + S(x_1) & K_f(x_1 - x_2) & \cdots & K_f(x_1 - x_k) \\ K_f(x_2 - x_1) & K_f(x_2 - x_2) + S(x_2) & \cdots & K_f(x_2 - x_k) \\ \vdots & \vdots & \vdots & \vdots \\ K_f(x_k - x_1) & K_f(x_k - x_2) & \cdots & K_f(x_k - x_k) + S(x_k) \end{bmatrix}_{k\times k}. \tag{7}$$

Из (4) можно выразить Q:

$$Q = K^{-1}Y. \tag{8}$$

Обратную матрицу K^{-1} представим следующим образом:

$$K^{-1} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,k} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{k,1} & v_{k,2} & \cdots & v_{k,k} \end{bmatrix}_{k\times k}. \tag{9}$$

Формулы (1), (2), (5) и (7) можно представить как некую модель, полученную после обработки данных.

Как видно из формулы (1), при большом количестве обучающих примеров решающие функции модели могут быть слишком громоздкими с точки зрения вычислительных затрат на их вычисление. Упрощение функций (1) зачастую является необходимым в практических задачах. Покажем, как математический аппарат метода позволяет быстрым способом выявлять слабые и шумовые элементы обучающей выборки, удаление которых поспособствует сокращению слагаемых в решающих функциях.

Выявление слабых и шумовых элементов выборки

Для возможности нахождения шумовых и слабых элементов выборки важно знать, как изменяются функции (1) при удалении некоторых элементов выборки.

Допустим, построена модель на всех обучающих данных, получены формулы (1), (2), (5) и (7). В статье [Бахвалов, Копылов, 2015] была установлена взаимосвязь между значением функции вида (1), полученной на выборке без x_i элемента, в точке x_i и реальным значением y_i, при существовании построенной модели на всей обучающей выборке. Данное соотношение легко переписать на случай многомерного выходного значения $y_i \in R^m$, тогда взаимосвязь можно записать следующим образом:

$$f_j^{(i)}(x_i) = y_{ij} - \frac{q_{i,j}}{v_{i,i}}, \tag{10}$$

где *i*-й индекс у функции $f_j^{(i)}$ означает, что функция (1) получена на выборке без x_i элемента; $q_{i,j}$ и $v_{i,i}$ — из (5) и (9).

Формула (10) позволяет быстро проводить процедуру оценки отклонений значений возможных функций от реальных значений. Подобная процедура дает возможность выявить элементы выборки, удаление которых слабо повлияет на вид результирующей функции. У подобных элементов отклонения $\frac{q_{i,j}}{v_{i,i}}$ обычно имеют небольшое значение. Также легко прослеживаются элементы, являющиеся пограничными с элементами других классов.

Рассмотрим синтетический пример. Дана выборка, состоящая из элементов трех классов: (x_i, y_i), где $x_i \in R^2$, $y_i \in R^3$, $i = 1,..,135$; построена модель, представленная в виде (1), (2), (5) и (7) (рис. 1, *а*, *б*).

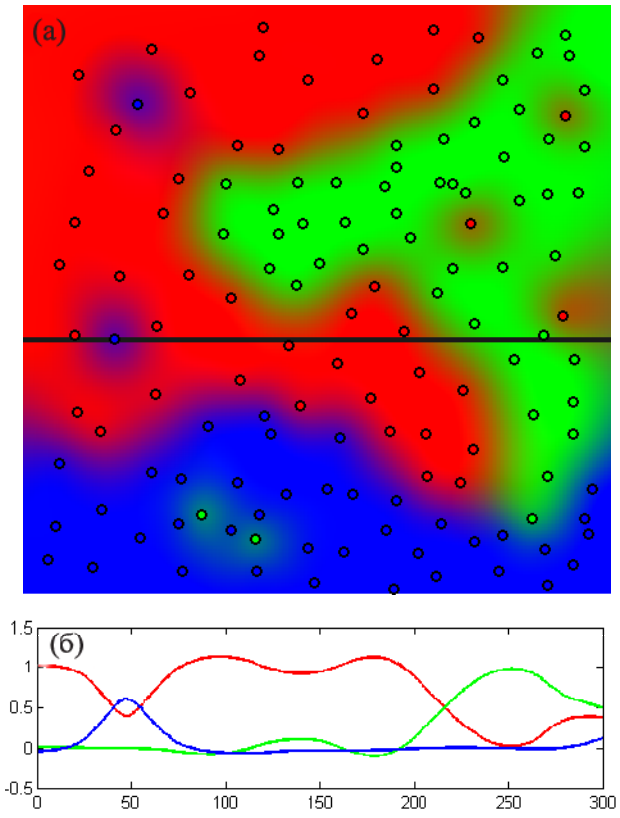

Рис. 1. (а) Графическое представление обученной модели классификатора. Модель обучена на 135 примерах, отмеченных кружками и залитыми цветом класса. По обеим осям отложены значения признаков элементов выборки. (б) Одномерный срез по одному из признаков, демонстрирующий виды полученных функций модели $f_1(x), f_2(x)$ и $f_3(x)$ при обучении. Срез соответствует черной линии (на рис. 1, *а*), по горизонтальной оси отложены значения одного из признаков элементов (второй признак фиксирован), по вертикальной оси отмечены значения построенных функций. Цветная версия рис. доступна на сайте журнала

Как видно из рис. 1, значения полученных решающих функций вида (1) изменяются приблизительно в диапазоне от 0 до 1. Причем чем ближе значение одной из функций к 1, тем ближе значения остальных функций к 0. Построив функции $f_1^*(x), f_2^*(x)$ и $f_3^*(x)$ без одного синего элемента x_{blue}, который лежит на срезе, получим вид функций, представленный на рис. 2.

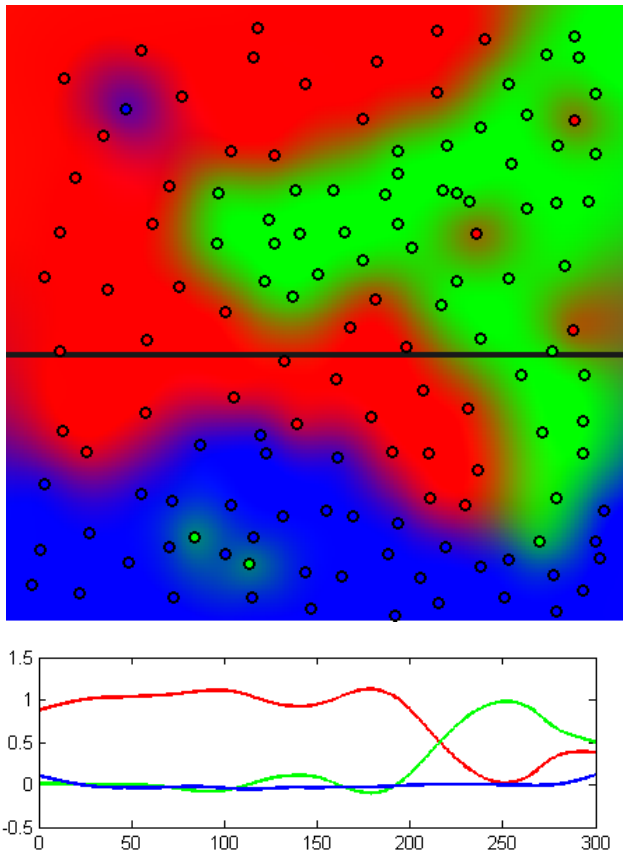

Рис. 2. Демонстрация изменения вида построенных функций (1) при удалении синего элемента, находящегося на срезе (черная линия). Цветная версия рис. доступна на сайте журнала

Из рис. 1 и 2 видно, как поменялся вид функций $f_1(x)$, $f_2(x)$ и $f_3(x)$ при удалении одного «сложного» элемента x_{blue} из выборки. В результате в удаленном элементе новая модель с функциями $f_1^*(x)$, $f_2^*(x)$ и $f_3^*(x)$ при классификации примет решение в пользу класса, обозначенного красным цветом.

Итак, значимость каждого (x_i, y_i) элемента выборки характеризуется величиной отклонения

$$e_i = \frac{q_{i,j}}{v_{i,i}}, \tag{11}$$

причем $y_{ij} = 1$.

Для выявления слабых или малоинформативных элементов выборки необходимо, чтобы величина e_i принимала отрицательные или близкие к 0 значения; у ошибочных или шумовых элементов значение e_i должно принимать значения, близкие к 1.

Экспериментальная часть

В качестве тестовых данных рассматривалась выборка полутоновых изображений (20 × 20 пикселей) цифр от 0 до 9, состоящая из 5000 элементов (рис. 3). Выборка была разделена пополам: на обучающую и тестовую.

Рис. 3. Пример изображений из выборки

Для возможности обучения классифицирующей модели каждое изображение было нормировано и представлено в виде вектора, т. е. в результате, обучающая выборка стала представима набором пар (x_i, y_i), где $x_i = (x_{i1}, x_{i2}, \ldots, x_{i400})$, $y_i = (y_{i1}, y_{i2}, \ldots, y_{i10})$. Модель, обученная на полной обучающей выборке, показала точность классификации 95.76 %.

Для выявления слабых и шумовых элементов были введены пороги *thWeak* и *thNoise*. Так, например, если для рассматриваемого *i*-го элемента значение $e_i \le thWeak$, то элемент считался малоинформативным и удалялся из обучающей выборки; также удаление происходило, если $e_i \ge thNoise$.

Представим графики зависимости числа найденных шумовых элементов и точности классификации модели, построенной без шумовых элементов, от значения порога *thNoise*. Значения порога *thNoise* изменялись в пределах от 0.5 до 1 с шагом 0.01 (рис. 4).

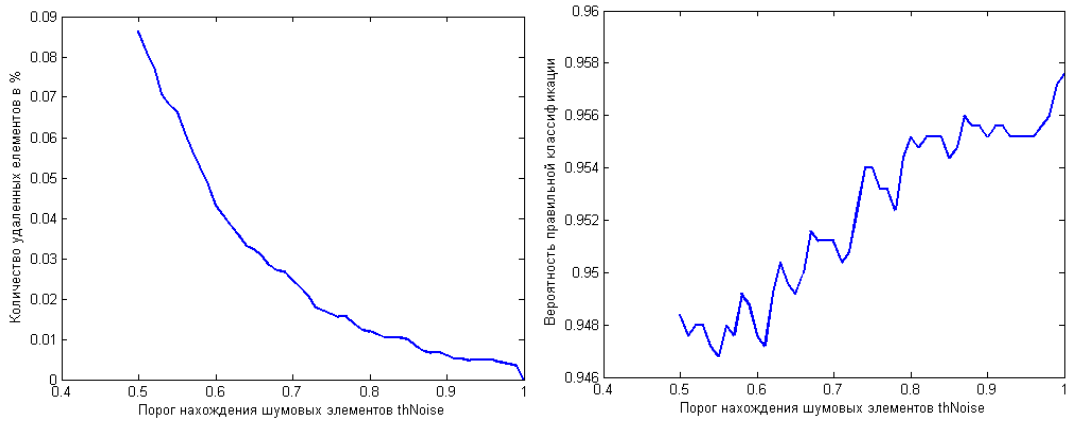

Рис. 4. Слева — график зависимости количества найденных шумовых элементов (в %-ном соотношении от общего числа обучающих элементов) от значений порога *thNoise*. Справа — график зависимости вероятности правильной классификации модели (обученной без найденных шумовых элементов) от значений порога *thNoise*

Представим графики зависимости числа найденных слабых элементов и точности классификации модели, построенной без них, от значения порога *thWeak*. Значения порога *thWeak* изменялись в пределах от 0 до 0.5 с шагом 0.01 (рис. 5).

Как видно из рис. 5, при $thWeak = 0.5$ было найдено и удалено более 91 % элементов обучающей выборки, которые были приняты за слабые или малоинформативные, при этом точность правильной классификации стала меньше 45 %, что не допустимо. С другой стороны, при $thWeak = 0.11$ из обучающей выборки было удалено более 50 % элементов, а точность при этом сократилась до 95.2 %; при этом количество коэффициентов в решающих правилах (1) сократилось более чем два раза.

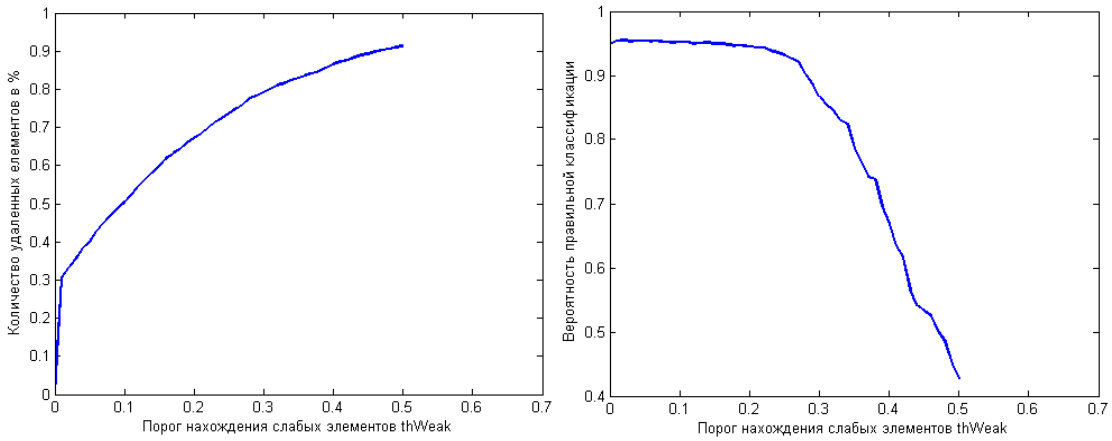

Рис. 5. Слева — график зависимости количества найденных слабых элементов (в %-ном соотношении от общего числа обучающих элементов) от значений порога *thWeak* . Справа — график зависимости вероятности правильной классификации модели (обученной без найденных слабых элементов) от значений порога *thWeak*

Для меньшего сокращения точности классификации возможно косвенное использование слабых элементов выборки при обучении; при этом количество слагаемых в решающих правилах остается таким же, как при обучении без уделенных элементов. В матрице K можно удалить столбцы, соответствующие слабым элементам, тогда формула (7) примет вид

$$K^* = \begin{bmatrix} K_f(x_1 - x_1) + S(x_1) & K_f(x_1 - x_2) & \cdots & K_f(x_1 - x_l) \\ K_f(x_2 - x_1) & K_f(x_2 - x_2) + S(x_2) & \cdots & K_f(x_2 - x_l) \\ \vdots & \vdots & \vdots & \vdots \\ K_f(x_k - x_1) & K_f(x_k - x_2) & \cdots & K_f(x_k - x_l) \end{bmatrix}_{k \times l}, \qquad (12)$$

где $l < k$. Для нахождения значений коэффициентов q_{ij} $(i = 1, \cdots, l; j = 1, \cdots, m)$ формула (8) изменится следующим образом:

$$Q = \left(K^T K\right)^{-1} K^T Y . \qquad (13)$$

Представим графики зависимости числа найденных слабых элементов и точности классификации модели, построенной с их косвенным использованием, от значения порога *thWeak* . Значения порога *thWeak* изменялись в пределах от 0 до 0.5 с шагом 0.01 (рис. 6).

При подходе с косвенным использованием слабых элементов при обучении видно, что результаты на тестах стали гораздо лучше: при удалении 91 % элементов выборки точность правильной классификации модели снизилась примерно на 5 % и составила 90 %; при сохранении точности классификации (95 %), близкой к исходной (95.76 %), возможно удалить гораздо больше элементов обучающей выборки — 72 %.

Такие результаты получаются на хорошо размеченной выборке. Если же в выборке присутствует большое количество ошибочно размеченных данных, то результаты классификации модели, обученной без найденных шумовых и слабых элементов, могут быть лучше, чем если бы использовалась вся обучающая выборка. Так, изменив у 15 % элементов выборки класс с правильного на ложный, модель, обученная на всей выборке, показала на тестовой выборке 90.6 % правильной классификации. При введенных порогах $thNoise = 0.9$ и $thWeak = 0.26$ из выборки было удалено 56 % элементов. Модель, обученная без шумовых элементов и с косвенным использованием слабых элементов, показала точность правильной классификации 94 % на тестовой выборке.

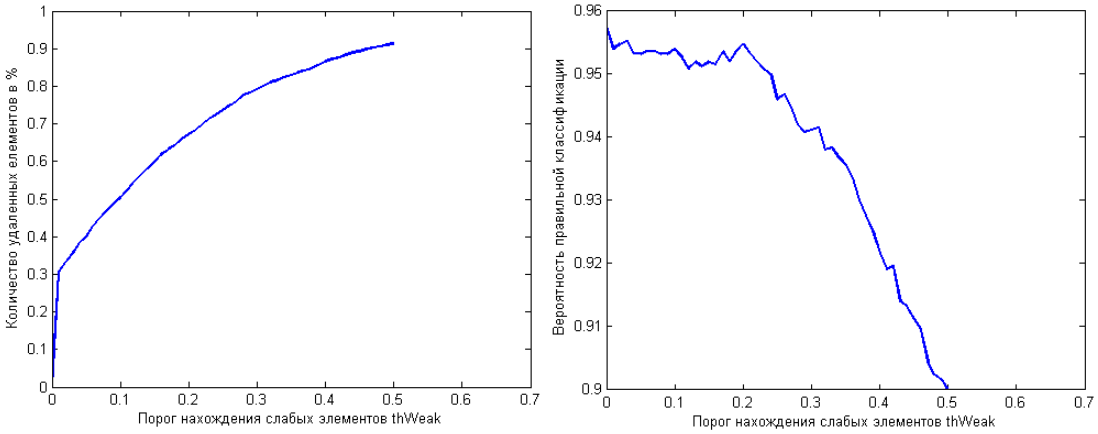

Рис. 6. Слева — график зависимости количества найденных слабых элементов (в %-ном соотношении от общего числа обучающих элементов) от значений порога *thWeak* . Справа — график зависимости вероятности правильной классификации модели (обученной с косвенным использованием слабых элементов) от значений порога *thWeak*

Исходную выборку цифр и выборку с ошибками разметки возможно скачать по ссылке [http://mallenom.ru/IvanK/DigitsData.zip]. Данные хранятся в .mat-файлах и представлены следующим образом: структура *data* , имеющая следующие поля *TeachX* , *TeachY* , *TestX* , *TestY* . *TeachX* и *TestX* представляют собой матрицы размером 2500×400, в которых хранятся нормализованные изображения, представленные в виде векторов строк. *TeachY* и *TestY* — вектор-столбцы, содержащие номера классов от 1 до 10, соответствующие цифрам (классу 10 соответствует цифра 0).

Заключение

В данной статье показано, что для метода машинного обучения на основе теории случайных функций существует возможность быстрого поиска слабых и шумовых элементов, удаление которых позволяет сократить вид решающих правил модели, тем самым ускорив ее работу. Кроме того, показано, что при наличии определенного процента ошибочно размеченных данных поиск и устранение ненужных элементов в итоге позволят улучшить обобщающие способности модели.

Список литературы (References)

Бахвалов Ю. Н., Копылов И. В. Обучение и оценка обобщающей способности методов многомерной интерполяции // Компьютерные исследования и моделирование. — 2015. — Т. 7, № 5. — С. 1023–1031.

Bahvalov Ju. N., Kopylov I. V. Obychenie i ocenka obobshaiyshej sposobnosti metodov mnogomernoj interpoliacii. [Training and assessment the generalization ability of multivariate interpolation method] // Computer Research and Modeling. — 2015. — Vol. 7, No. 5. — P. 1023–1031 (in Russian).

Бахвалов Ю. Н., Зуев А. Н., Ширабакина Т. А. Метод распознавания образов на основе теории случайных функций // Санкт Петербург: Известия вузов. Приборостроение. — 2005. — Т. 48, № 2. — С. 5–8.

Bahvalov Ju. N., Zuev A. N., Shirabakina T. A. Metod raspoznavanija obrazov na osnove teorii sluchajnyh funkcij [The method of pattern recognition based on the theory of random functions] // St. Peterburg: Izvestija vuzov. Priborostroenie. — 2005. — Vol. 48, No. 2. — P. 5–8 (in Russian).

Бахвалов Ю. Н., Малыгин Л. Л., Черкас П. С. Метод машинного обучения на основе алгоритма многомерной интерполяции и аппроксимации случайных функций // Вестник Череповецкого государственного университета. — 2012. — Т. 2, № 2. — С. 7–9.

Bahvalov Ju. N., Malygin L. L., Cherkas P. S. Metod mashinnogo obuchenija na osnove algoritma mnogomernoj interpoljacii i approksimacii sluchajnyh funkcij [The method of machine learning based on algorithm of multidimensional interpolation and approximation of random functions] // Vestnik Cherepovetskogo gosudarstvennogo universiteta. — 2012. — Vol. 2, No. 2. — P. 7–9 (in Russian).

Выборка [Электронный ресурс]: http://mallenom.ru/IvanK/DigitsData.zip
Training dataset [Electronic resource]: http://mallenom.ru/IvanK/DigitsData.zip

Кошель С. М., Мусин О. Р. Методы цифрового моделирования: кригинг и радиальная интерполяция // Информационный бюллетень ГИС-Ассоциации. — 2000. — №4(26)–5(27). — С. 32–33.
Koshel S. M., Musin O. R. Metody cifrovogo modelirovanija: kriging i radial'naja interpoljacija [Digital simulation methods: kriging and radial interpolation] // Informacionnyj bjulleten' GIS-Associacii. — 2000. — No. 4(26)–5(27). — P. 32–33 (in Russian).

Robeson S. M. Spherical Methods for Spatial Interpolation: Review and Evaluation // Cartography and Geographic Information Systems. — 1997. — Vol. 24, No.1. — P. 3–20.