Виконано побудову довірчого інтервалу нелінійного рівняння регресії часу відновлення працездатності пристроїв термінальної мережі на основі нормалізуючого перетворення Джонсона та t-розподілу Стьюдента, без припущення про нормальність емпіричних даних. Проведено порівняння отриманих результатів з результатами, отриманими іншими методами. Перетворення Джонсона в порівнянні з іншими нормалізуючими перетвореннями дозволяє краще здійснити нормалізацію емпіричних даних

Ключові слова: довірчий інтервал, нелінійна регресія, нормалізуюче перетворення, перетворення Джонсона, термінальна мережа

Выполнено построение доверительного интервала нелинейного уравнения регрессии времени восстановления работоспособности устройств терминальной сети на основе нормализирующего преобразования Джонсона и t-распределения Стьюдента, без предположения о нормальности эмпирических данных. Проведено сравнение полученных результатов с результатами, полученными другими методами. Преобразование Джонсона по сравнению с другими нормализирующими преобразованиями позволяет лучше осуществить нормализацию эмпирических данных

Ключевые слова: доверительный интервал, нелинейная регрессия, нормализирующее преобразование, преобразование Джонсона, терминальная сеть

УДК 004.942:519.25

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ НЕЛИНЕЙНОЙ РЕГРЕССИИ ВРЕМЕНИ ВОССТАНОВЛЕНИЯ РАБОТОСПОСОБНОСТИ УСТРОЙСТВ ТЕРМИНАЛЬНОЙ СЕТИ

С. Б. Приходько
Доктор технических наук, доцент*
E-mail: sergiy.prykhodko@nuos.edu.ua
Л. Н. Макарова
Соискатель*
E-mail: lidiya@ultra.mk.ua
*Кафедра программного обеспечения автоматизированных систем
Национальный университет кораблестроения им. Адмирала Макарова
пр. Героев Сталинграда, 9, г. Николаев, Украина, 54025

1. Введение

Повышение надежности оценки времени восстановления работоспособности устройств терминальной сети играет важную роль в практических задачах ее управления [1]. Как правило, время восстановления работоспособности устройств терминальной сети является негауссовской случайной величиной (СВ), которая зависит от ряда факторов, в том числе от расстояния между центром обслуживания и устройством терминальной сети, вида отказа в обслуживании, модели конкретного устройства и т.д. Для оценки времени восстановления работоспособности устройств терминальной сети необходимо построить соответствующую регрессионную модель [2, 3], которая будет являться нелинейной [4]. Повысить надежность ее оценки можно за счет построения доверительного интервала нелинейной регрессии [5, 6].

В случае негауссовской СВ построение доверительного интервала нелинейной регрессии без предположения о нормальности СВ затруднено. Применение такого предположения может существенно исказить результаты.

Поэтому проблема построения доверительного интервала нелинейной регрессии времени восстановления работоспособности устройств терминальной сети является актуальной.

2. Цель исследования

Целью исследования является построение доверительного интервала нелинейной регрессии времени восстановления работоспособности устройств терминальной сети без предположения о нормальности СВ.

Для достижения поставленной цели исследования необходимо решить следующие задачи:

1. Нормализировать эмпирические данные, для чего осуществить выбор нормализирующего преобразования и оценить его параметры.

2. На основе нормализированных данных получить линейное уравнение регрессии и построить для него доверительный интервал.

3. Используя выбранное нормализирующее преобразование, построенные линейную регрессию и ее доверительный интервал, осуществить переход к нелинейной регрессии и ее доверительному интервалу.

3. Литературный обзор

При нормальном законе распределения СВ доверительный интервал линейного уравнения регрессии возможно построить традиционным методом с использованием t-распределения Стьюдента [7]. Однако для нелинейной регрессии данный метод не учитывает ряд особенностей эмпирического распределения данных, например его асимметрию.

Использование линеаризирующих преобразований сводится к получению линейной регрессионной модели из исходной нелинейной путем замены переменных и коэффициентов.

Однако такая замена приводит к упрощению регрессионной модели и некоторой потере информации, связанной с нелинейностью [8–10].

Применение нормализирующих преобразований позволяет перейти к линейной регрессии для нормализированных данных, для нее построить доверительный интервал традиционным способом с использованием t-распределения Стьюдента, а затем путем применения соответствующего преобразования перейти к нелинейной регрессии и ее доверительному интервалу [11–13]. Данный подход лишен недостатков, отмеченных у предыдущих методов.

В качестве нормализирующего преобразования используется преобразование Джонсона, так как в ряде случаев оно дает лучшие результаты по сравнению с другими известными преобразованиями, например, Бокса-Кокса (Box-Cox) [11, 14, 15].

4. Построение доверительного интервала нелинейной регрессии

В общем виде регрессионная модель может быть представлена следующим уравнением [3]:

$$y = \bar{y} + \varepsilon_t = f(x) + \varepsilon_t, \qquad (1)$$

где y – зависимая переменная, или результативный признак, x – независимая переменная, или фактор, ε_t – случайная ошибка, или возмущение, f(x) – функция, которая определяет вид регрессионной модели: нелинейная или линейная.

Доверительный интервал линейной регрессии возможно построить с помощью t-распределения Стьюдента по следующей формуле [7]:

$$y = \hat{y} \pm t_{(\alpha/2,n-2)} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \qquad (2)$$

где $t_{(\alpha/2,n-2)}$ – квантиль t-распределения Стьюдента, $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$, α – доверительная вероятность, n – количество значений СВ в выборке, $\hat{y}$ – значение y, рассчитанное по уравнению регрессии.

В случае нелинейной регрессионной модели необходимо перейти к линейной с помощью линеаризирующего или нормализирующего преобразований. Далее для полученной линейной регрессионной модели построить доверительный интервал уравнения регрессии по следующей формуле [9]:

$$y = \hat{y} \pm u(\nu,\beta,\lambda) \cdot S \cdot \sqrt{\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{1}{n}}, \qquad (3)$$

где u(υ, β, λ) – табличный коэффициент, зависящий от υ, β, λ, $\upsilon = n-2$, $\beta = 1-\alpha$, $\lambda = \sqrt{\frac{1}{2}\cdot\left(1 - \frac{1+nCD}{(1+nC^2)(1+nD^2)}\right)}$,

$C = \frac{x_1-\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$, $D = \frac{x_2-\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$, $S = \sqrt{\frac{1}{\nu}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$,

n – количество значений в выборке, α – доверительная вероятность, x_1, x_2 – границы заданного отрезка.

При использовании нормализирующего преобразования Джонсона для получения нелинейной регрессионной модели необходимо нормализировать исходные СВ с помощью преобразования, которое в общем виде представлено следующей формулой [16]:

$$z = \gamma + \eta h(x,\phi,\lambda);\ \eta > 0;\ -\infty < \gamma < \infty;\ \lambda > 0;\quad -\infty < \phi < \infty. \qquad (4)$$

Преобразование (4) имеет обратное преобразование:

$$x = \phi + \lambda h^{-1}(z,\gamma,\eta);\ \eta > 0;\ -\infty < \gamma < \infty;\ \lambda > 0;\ -\infty < \phi < \infty, \qquad (5)$$

где z – нормально распределенная СВ с математическим ожиданием ноль и дисперсией единица; x – СВ с распределением Джонсона; γ, η, φ, λ – параметры преобразования или распределения Джонсона; h и h^{-1} – функции определенного семейства:

$$h = \begin{cases} \ln(\tilde{x}), & x > \phi, \quad \text{для семейства } S_L; \\ \ln\left[\tilde{x}/(1-\tilde{x})\right], & \phi < x < \phi + \lambda, \quad \text{для семейства } S_B; \\ \mathrm{Arsh}(\tilde{x}), & -\infty \le x \le +\infty, \quad \text{для семейства } S_U, \end{cases}$$

$$h^{-1} = \begin{cases} e^{\zeta}, & \text{для семейства } S_L; \\ 1/\left(1+e^{-\zeta}\right), & \text{для семейства } S_B; \\ \left(e^{\zeta}-e^{-\zeta}\right)/2, & \text{для семейства } S_U. \end{cases}$$

Конкретное семейство распределений Джонсона выбирается исходя из значений квадрата асимметрии A^2 и эксцесса ε исходной выборки [17]. Значения неизвестных параметров распределения можно найти с помощью непараметрического метода решения задачи математического программирования, описанного в [11]:

$$\hat{\theta} = \arg\min_{\theta}\left\{\hat{A}_z^2 + (\hat{\varepsilon}_z - 3)^2 + \hat{m}_z^2 + (\hat{D}_z - 1)^2\right\}, \qquad (6)$$

где θ – вектор неизвестных параметров, θ={γ, η, φ, λ}, $\hat{A}_z^2$ – оценка асимметрии распределения z, $\hat{\varepsilon}_z$ – оценка эксцесса распределения z, $\hat{m}_z^2$ – оценка математического ожидания z, $\hat{D}_z$ – оценка дисперсии z, n – количество значений z в выборке.

Проверку соответствия преобразованных выборок нормальному распределению можно выполнить с помощью критериев согласия, например, χ^2 Пирсона или Колмогорова-Смирнова [18].

В общем виде линейная регрессионная модель нормализированных значений СВ может быть представлена уравнением:

$$z_y = b_1 z_x + b_0, \quad (7)$$

где b_1, b_0 – коэффициенты линейной регрессии, которые находятся методом наименьших квадратов.

Для проверки адекватности линейной регрессионной модели используем коэффициент детерминации R^2 [19]:

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2 \right), \quad (8)$$

где y_i – фактическое значение y, $\hat{y}_i$ – расчетное значение y, $\bar{y}$ – среднее значение y.

Чем ближе значение R^2 к единице, тем выше качество модели: приемлемой считается модель при $R^2 \geq 0{,}5$, достаточно качественной – при $R^2 \geq 0{,}8$, при $R^2 = 1$ линия регрессии точно соответствует всем наблюдениям.

Алгоритм построения (1-α) % доверительного интервала линейного уравнения регрессии можно представить следующим образом.

1. При каждом фиксированном значении z_{xi} рассмотрим соответствующие ему значения z_{yi}, $i \in [0; m]$, как малую выборку, которую повторно нормализируем с помощью преобразования Джонсона. Для полученной выборки z_y^* найдем выборочное среднее m_{zy}^* и выборочное среднеквадратическое отклонение S_{zy}^*.

Проверку соответствия преобразованных выборок нормальному распределению можно выполнить с помощью критерия согласия Колмогорова-Смирнова по формулам, приведенным в [18].

2. Используя традиционный способ на основе t-распределения Стьюдента для выборочного среднего нормальной выборки [20] найдем $\Delta z_y^* = t_{n-1} S_{zy}^* / \sqrt{m}$ и получим границы доверительного интервала выборочного среднего СВ z_y^*: $z_y^* \min = \bar{z}_y^* - \Delta zy^*$ и $z_y^* \max = \bar{z}_y^* + \Delta zy^*$.

3. По обратному преобразованию (5) для значений $z_y^* \min$ и $z_y^* \max$ получим границы доверительного интервала выборочного среднего СВ z_y: $z_y \min$ и $z_y \max$.

4. По всем значениям $z_y \min$ и $z_y \max$ построим нижнюю и верхнюю границы доверительного интервала уравнения регрессии с помощью квадратичного полинома вида:

$$z_y = d_2 z_x^2 + d_1 z_x + d_0, \quad (9)$$

где d_2, d_1, d_0 – коэффициенты уравнения (9), которые находим методом наименьших квадратов.

Используя линейную регрессионную модель (7) и нормализирующее преобразование Джонсона (4), построим нелинейную регрессионную модель времени восстановления работоспособности устройств терминальной сети:

$$y = \frac{e^{c_1}(\lambda_y + \phi_y) + \phi_y}{1 + e^{c_1}}, \quad (10)$$

где $c_1 = \frac{1}{\eta_y} \cdot \left(b_1 \cdot z_x + b_0 - \gamma_y \right)$, $z_x = \gamma_x + \eta_x \ln\left(\frac{x - \phi_x}{\lambda_x + \phi_x - x} \right)$.

(1-α) % доверительный интервал нелинейного уравнения регрессии можно построить, используя линейную регрессионную модель (7), t-распределение Стьюдента(2),инормализирующеепреобразованиеДжонсона (4):

$$y = \frac{e^{k_1}(\lambda_y + \phi_y) + \phi_y}{1 + e^{k_1}}, \quad (11)$$

где

$$k_1 = \frac{1}{\eta_y} \cdot \left(b_1 \cdot z_x + b_0 - \gamma_y \pm t_{(\alpha/2, n-2)} \cdot S_{z_y} \cdot \sqrt{\frac{1}{n} + \frac{(z_x - \bar{z}_x)^2}{\sum_{i=1}^{n} (z_{xi} - \bar{z}_x)^2}} \right),$$

$$z_x = \gamma_x + \eta_x \ln\left(\frac{x - \phi_x}{\lambda_x + \phi_x - x} \right).$$

Формула (11) фактически является уравнением верхней и нижней границ доверительного интервала и позволяет получать несимметричный доверительный интервал нелинейного уравнения регрессии.

5. Практическая реализация предложенного метода построения доверительного интервала нелинейной регрессии

Практическая реализация результатов исследования была выполнена на основании данных, приведенных в [15]. Расчет велся на примере одной из выборок данных, где СВ x – расстояние от центра обслуживания до терминального устройства, м; СВ y – время восстановления работоспособности терминального устройства, мин. СВ x и СВ y не являются гауссовскими: A_x=1,0885; ε_x=2,9193; A_y=1,6538; ε_y=5,2015. С помощью нормализирующего преобразования Джонсона (4) выполним нормализацию СВ x и СВ y. Исходя из значений квадрата асимметрии и эксцесса для нормализации СВ x и СВ y было выбрано семейство распределений S_B Джонсона.

Параметры преобразования найдем в результате решения задачи (6). Для СВ x: γ_x=0,9676; η_x=0,5755; φ_x=1874,59; λ_x=8300,00. Для СВ y: γ_y=1,2500; η_y=0,5431; φ_y=10,2939; λ_y=2360,61. С доверительной вероятностью 0,95 гипотеза о соответствии преобразованных выборок нормальному закону распределения СВ принимается: для СВ z_x значение χ^2_{zx}=6,25 при критическом значении χ^2=11,07, для СВ z_y значение χ^2_{zy}=6,14 при критическом значении χ^2=9,49, при оценке χ^2 для СВ z_y было произведено объединение двух подинтервалов ввиду малого количества значений (n_j<5) в одном из них.

В соответствии с (7) найдем линейное уравнение регрессии нормализированных значений: $z_y = 0{,}8765 z_x + 0{,}0495$, коэффициент детерминации R^2, вычисленный по формуле (8), составляет 0,9778.

Для найденного линейного уравнения регрессии построим 95 % доверительный интервал по формуле (2), уравнения границ которого имеют вид $z_y = 0{,}8765 z_x + 0{,}0495 \pm 1{,}1405 \sqrt{0{,}0070 + \frac{(z_x + 0{,}0578)^2}{142{,}78}}$.
Для уравнений обеих границ коэффициент детерминации R^2 составляет 0,7452, что говорит об адекватности найденных уравнений границ.

Для того же линейного уравнения регрессии построим 95 % доверительный интервал по формуле (3), уравнения границ которого имеют вид $z_y = 0{,}8765 z_x + 0{,}0495 \pm 1{,}1502 \sqrt{\frac{(z_x + 0{,}0578)^2}{142{,}78} + 0{,}0070}$.
Для уравнений обеих границ коэффициент детерминации R^2 составляет 0,7449, что также говорит об адекватности найденных уравнений границ, однако полученный результат несколько хуже, чем результат, полученный по формуле (2).

По приведенному выше алгоритму построим 95 % доверительный интервал линейной регрессии в соответствии с (9). 95 % доверительный интервал линейного уравнения регрессии нормализированного времени восстановления работоспособности устройств терминальной сети приведен на рис. 1.

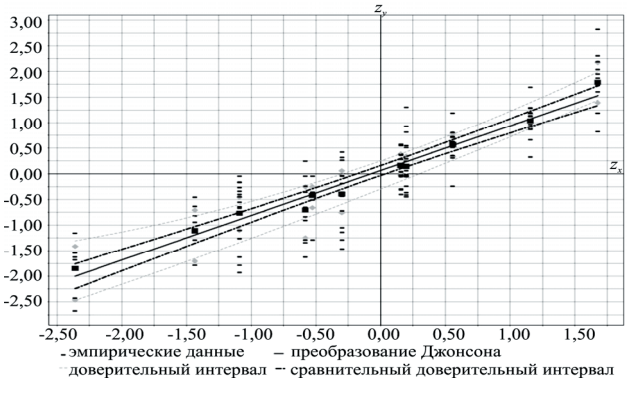


Рис. 1. 95 % доверительный интервал линейного уравнения регрессии нормализированного времени восстановления работоспособности устройств терминальной сети

Уравнение нижней границы имеет вид $z_y = 0{,}0293 z_x^2 + 0{,}988 z_x - 0{,}2975$, коэффициент детерминации R^2 составляет 0,978. Уравнение верхней границы имеет вид $z_y = 0{,}0936 z_x^2 + 0{,}88 z_x + 0{,}25$, коэффициент детерминации R^2 составляет 0,9848. Полученные значения R^2 говорят о достаточно качественных моделях построенных уравнений границ доверительного интервала. Кроме того, данный результат существенно лучше, чем результаты, полученные по формулам (2) и (3).

При построении доверительного интервала линейного уравнения регрессии по формулам (2) и (3) он оказывается симметричным относительно самого уравнения, а это не совсем адекватно описывает эмпирическое распределение данных. Применение преобразования Джонсона для построения доверительного интервала линейного уравнения регрессии дает более адекватное представление эмпирических данных, а именно:

– при меньшем количестве точек и большей дисперсии данных доверительный интервал расширяется, что можно наблюдать ближе к границам отрезка построения доверительного интервала;

– при большем количестве точек и меньшей дисперсии данных доверительный интервал сужается, что можно наблюдать в середине отрезка построения доверительного интервала.

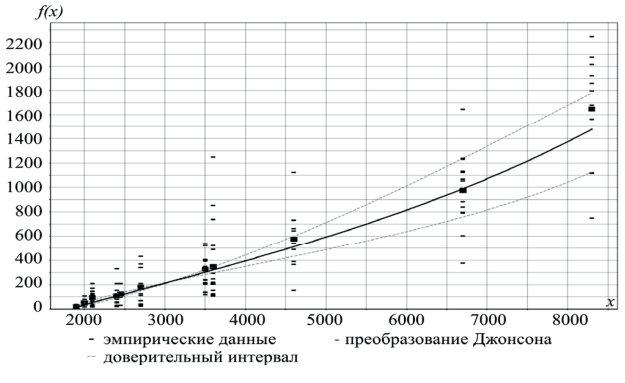
В соответствии с (10) найдем нелинейное уравнение регрессии времени восстановления работоспособности устройств терминальной сети: $y = \frac{2370{,}90 \cdot e^{0{,}9288 \ln\left(\frac{x-1874{,}59}{10174{,}59-x}\right) - 0{,}6488} + 10{,}29}{1 + e^{0{,}9288 \ln\left(\frac{x-1874{,}59}{10174{,}59-x}\right) - 0{,}6488}}$, коэффициент детерминации $R^2 = 0{,}8082$.

В соответствии с (11) построим 95 % доверительный интервал нелинейного уравнения регрессии времени восстановления работоспособности устройств терминальной сети (рис. 2, *а*), границы которого имеют вид $y = \frac{2370{,}90 \cdot e^k + 10{,}29}{1 + e^k}$,

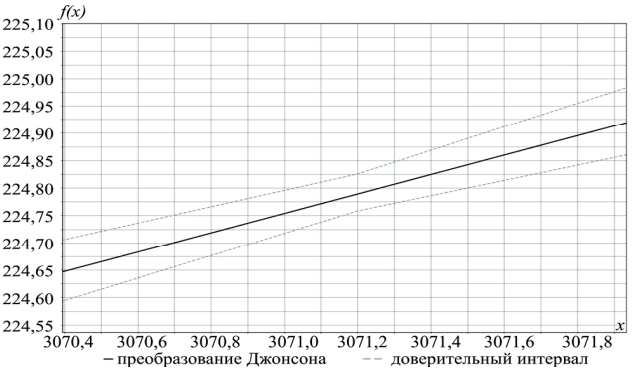
где

$$k = 0{,}9288 z - 0{,}6488 \pm 4{,}1704 \sqrt{0{,}0023 z^2 + 0{,}0083 z + 0{,}0144},$$

$$z = \ln\left(\frac{x - 1874{,}59}{10174{,}59 - x}\right).$$



а



б

Рис. 2. 95 % доверительный интервал нелинейного уравнения регрессии времени восстановления работоспособности устройств терминальной сети: *а* – вся область; *б* – увеличенный фрагмент части области

На рис. 2, *б* приведен увеличенный фрагмент отрезка построения доверительного интервала. Из этого рисунка видно, что границы доверительного интервала не пересекаются, а достаточно близко подходят к графику уравнения регрессии. Таким образом, доверительный интервал на этом участке значительно уже по сравнению с остальным отрезком построения.

Применение нормализирующего преобразования Джонсона семейства S_B при построении доверительного интервала нелинейной регрессии более адекватно описывает распределение эмпирических данных по сравнению с формулами (2) или (3):

– доверительный интервал не симметричный относительно регрессии в случае не нулевой асимметрии распределения эмпирических данных;

– при большем количестве данных и меньшей их дисперсии доверительный интервал значительно уже;

– при меньшем количестве данных и большей их дисперсии доверительный интервал значительно шире.

6. Выводы

Предложен способ построения доверительного интервала нелинейного уравнения регрессии времени восстановления работоспособности устройств терминальной сети на основе нормализирующего преобразования Джонсона семейства S_B без предположения о нормальности СВ. Суть способа состоит в следующем. Сначала необходимо нормализировать эмпирические данные, для чего осуществить выбор нормализирующего преобразования и оценить его параметры. Далее на основе нормализированных данных получить линейное уравнение регрессии и построить для него доверительный интервал. И, наконец, используя выбранное нормализирующее преобразование, построенные линейную регрессию и ее доверительный интервал, перейти к нелинейной регрессии и ее доверительному интервалу.

В отличие от существующих, предложенный способ построения доверительного интервала нелинейной регрессии не требует принятия предположения о нормальности СВ и позволяет учитывать ряд особенностей эмпирического распределения данных, например, его реальную асимметрию и эксцесс. Применение преобразования Джонсона по сравнению с другими известными нормализирующими преобразованиями, например, Бокса-Кокса (Box-Cox), обусловлено и тем, что лишь для преобразования Джонсона по значениям асимметрии в квадрате и эксцесса можно заранее выбрать соответствующее его семейство, с помощью которого удастся осуществить нормализацию данных. Кроме того, преобразование Джонсона семейства S_B подходит для нормализации данных с бимодальными и U-образными распределениями. Однако вместе с тем с помощью преобразования Джонсона семейства S_B (в отличие от S_U) получают не изоморфные множества значений, что обуславливает приближенность нормализации.

Однако подобное ограничение, на наш взгляд, на практике можно не принимать во внимание, так как выборка реальных данных всегда ограничена. Вместе с тем в дальнейшем, если удастся найти соответствующие данные, предложенный способ построения доверительного интервала нелинейной регрессии предполагается модифицировать для преобразования Джонсона семейства S_U.

Литература

1. Наша цель – банк в шаговой доступности [Электронный ресурс] / Режим доступа: http://www.inpas.ru/publications/78/ – Загл. с экрана.
2. Грешилов, А. А. Математические методы построения прогнозов [Текст] / А. А. Грешилов, В. А. Стакун, А. А. Стакун. – М.: Радио и связь, 1997. – 112 с.
3. Демиденко, Е.З . Линейная и нелинейная регрессии [Текст] / Е. З. Демиденко. – М.: Финансы и статистика, 1981. – 302 с.
4. Bates, Douglas M. Nonlinear Regression Analysis and Its Applications [Text] / Douglas M. Bates, Donald G. Watts. – Wiley, 1988. – 384 p.
5. Pardoe, Iain Applied regression modeling [Text] / Iain Pardoe. – Wiley, 2012. – 325 p.
6. Seber, George A. F. Nonlinear Regression [Text] / George A. F. Seber, C. J. Wild. – John Wiley & Sons, Inc., 2003. – 792 p.
7. Yan, Xin Linear regression analysis: theory and computing [Text] / Xin Yan, Xiao Gang Su. – Singapore: World Scientific Publishing Co. Pte. Ltd., 2009. – 328 p.
8. Айвазян, С. А. Прикладная статистика. Основы эконометрики: Учебник для вузов [Текст]: В 2 т. 2-е изд., испр. – Т. 1: Теория вероятностей и прикладная статистика / С. А. Айвазян, В. С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001. – 656 с.
9. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников [Текст] / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.
10. Chatterjee, Samprit Handbook of Regression Analysis [Text] / Samprit Chatterjee, Jeffrey S. Simonoff. – Wiley, 2012. – 240 p.
11. Приходько, С. Б. Інтервальне оцінювання статистичних моментів негаусівських випадкових величин на основі нормалізуючих перетворень [Текст] / С. Б. Приходько // Математичне моделювання: науковий журнал. – Дніпродзержинськ: ДДТУ. – 2011. – № 1 (24). – С. 9–13.
12. Приходько, С. Б. Метод побудови нелінійних рівнянь регресії на основі нормалізуючих перетворень [Текст] : тези доп. міждерж. наук.-методич. конф. / С. Б. Приходько // Проблеми математичного моделювання. – Дніпродзержинськ: ДДТУ, 2012. – С. 31–33.
13. Ryan, Thomas P. Modern Regression Methods [Text] / Thomas P. Ryan – Wiley, 2008. – 672 p.

14. Приходько, С. Б. Розробка нелінійної регресійної моделі тривалості програмних проектів на основі нормалізуючого перетворення Джонсона [Текст] / С. Б. Приходько, А. В. Пухалевич // Радіоелектронні і комп`ютерні системи. – – 2012. – № 4 (56). – С. 90–93.
15. Приходько, С. Б. Определение доверительных интервалов статистических моментов времени наработки между отказами устройств терминальной сети [Текст] / С. Б. Приходько, Л. Н. Макарова // Наукові праці: науково-методичний журнал. Комп`ютерні технології. – 2013. – Вип. 201, Т. 213. – С. 82–86.
16. Кендалл, М. Теория распределений [Текст] / М. Кендалл, А. Стьюарт. – М.: Наука, 1966. – 588 с.
17. Johnson, N. L. System of Frequency Curves Generated by Methods of Translation [Text] / N. L. Johnson // Biometrica. – 1949. – Vol. 36, № ½. – P. 149–176.
18. Вентцель, Е. С. Теория вероятностей: Учеб. для вузов [Текст] / Е. С. Вентцель. – М.: Высш. шк., 1999. – 576 с.
19. Магнус, Я. Р. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. [Текст] / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М.: Дело, 2004. – 576 с.
20. Поллард, Дж. Справочник по вычислительным методам статистики [Текст] / Дж. Поллард; пер. с англ. В. С. Занадворова; под ред. и с предисл. Е. М. Четыркина. – М.: Финансы и статистика, 1982. – 344 с.

Запропоновано новий метод видалення викидів з навчальних вибірок систем розпізнавання, заснований на побудові скорочених зважених вибірок w-об'єктів. Запропоновано алгоритми видалення викидів при порогах фільтрації, що визначаються автоматично та встановлюються користувачем. Наведено результати експериментальних досліджень, що підтверджують ефективність запропонованого методу

Ключові слова: навчаюча вибірка, фільтрація даних, викид, w-об'єкт, вирішуюче правило, утворююча множина

Предложен новый метод удаления выбросов из обучающих выборок систем распознавания, основанный на построении сокращенных взвешенных выборок w-объектов. Предложены алгоритмы удаления выбросов при автоматическом и определяемом пользователем порогах фильтрации. Приведены результаты экспериментальных исследований, подтвердивших эффективность предложенного метода

Ключевые слова: обучающая выборка, фильтрация данных, выброс, w-объект, решающее правило, образующее множество

УДК 004.67

МЕТОД УДАЛЕНИЯ ВЫБРОСОВ В ДАННЫХ НА ОСНОВЕ ВЗВЕШЕННЫХ ОБУЧАЮЩИХ ВЫБОРОК W-ОБЪЕКТОВ

Е. В. Волченко
Кандидат технических наук, доцент
Кафедра программного обеспечения
интеллектуальных систем
«Донецкий национальный
технический университет»
пр. Б. Хмельницкого, 84, г. Донецк,
Украина, 83050
E-mail: LenaVLV@gmail.com

1. Введение

Проблема качества данных является на сегодняшний день одной из важнейших проблем, решаемых при построении интеллектуальных систем [1–3]. Особенно остро данная проблема проявляется при построении обучающихся систем распознавания как самостоятельных систем или подсистем сложных интеллектуальных систем [4].

От качества исходных данных зависит не только достоверность функционирования системы в будущем, но и возможность её обучения в принципе [1], поэтому для повышения качества эмпирических данных в большинстве систем выполняется их предварительная обработка [2, 3].

Предобработка данных в системах распознавания является итеративным процессом и включает [1]:

– очистку данных, которая заключается в удалении шума, пропусков в данных и данных низкого качества;

– сжатие данных, включающее в себя нахождение минимального пространства признаков и репрезентативного множества данных на основе методов редукции и трансформации;