ISSN 1814-1196
Научный вестник НГТУ
том 69, № 4, 2017, с. 105–116

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 69, No. 4, 2017, pp. 105–116

ОБРАБОТКА ИНФОРМАЦИИ INFORMATION PROCESSING

УДК 519.816

Построение модели бинарного выбора на основе устойчивого семейства распределений*

В.С. ТИМОФЕЕВ[1], А.А. САНИНА[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, доцент. E-mail: v.timofeev@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант. E-mail: anastas.sanina@gmail.com*

В данной статье рассмотрена проблема, которая наиболее часто встречается при решении задач классификации, – повышение качества модели с точки зрения количества верно классифицированных объектов. Наиболее известные модели (логит- и пробит-), применяемые при решении задач классификации, показывают хорошие результаты далеко не на всех наборах данных. В связи с этим предпринимается попытка улучшить качество классификации путем выбора другого универсального семейства распределений, наилучшим способом описывающего реальные данные, вместо логистического или нормального законов распределений для построения модели. В работе рассмотрен пример решения задачи классификации для двух входных переменных с применением модели бинарного выбора, построенной на основе логистического, а затем устойчивого семейства распределений. Выполнены сравнение и оценка полученных решений, основанные на вычислении ошибки классификации. Предложен усовершенствованный подход для решения задачи классификации. Его основное отличие состоит не только в использовании «нового» семейства распределения для построения модели бинарного выбора, но и в дополнительной процедуре оптимизации, в ходе которой происходит варьирование параметров устойчивого семейства распределения с целью уменьшения ошибки классификации. Для детального исследования данного подхода с использованием технологии статистического моделирования проведен ряд вычислительных экспериментов. Полученные результаты свидетельствуют об эффективности предложенного метода. Выполнено сравнительное исследование качества классификации для различных моделей при разной степени засорения исходных данных. Выявлен характер зашумления данных, оказывающий существенное влияние на качество классификации. Указаны возможные перспективы развития работы. Важно отметить, что предложенный метод имеет существенно лучшее качество классификации, особенно на данных с определенным типом зашумления, а значит, может быть рекомендован для применения на практике.

Ключевые слова: задача классификации, логит-модель, модель бинарного выбора, устойчивое распределение, ошибка классификации, функция правдоподобия, выбросы, зашумленные данные, характеристическая функция, вычислительный эксперимент, оценка параметров, функция распределения

DOI: 10.17212/1814-1196-2017-4-105-116

---

* *Статья получена 27 октября 2017 г.*

ВВЕДЕНИЕ

В современном мире каждый человек периодически сталкивается с проблемой выбора. Ограниченность ресурсов или необходимость заставляют человека принимать то или иное решение. Всегда хочется, чтобы сделанный шаг оказался наиболее правильным и выгодным. Современная наука уже может помочь с решением таких трудностей. На языке статистики данный класс задач описывается как задачи классификации, для решения которых существуют различные методы дискриминантного анализа. Частным случаем таких подходов является метод построения логит- и пробит-моделей. В основе этих моделей лежат логистическое и нормальное распределения соответственно [1, 2].

Упомянутые модели эффективны с точки зрения качества классификации, если входные данные представляют собой непрерывные и независимые переменные, а также несущественно засорены ошибочно классифицированными наблюдениями. Описанные ограничения значительно сокращают круг задач, для решения которых могут быть использованы логит- и пробит-модели [10, 11].

В связи с тем, что множество реальных данных зачастую не соответствует всем указанным требованиям для применения логит- и пробит-моделей, качество классификации в этом случае оставляет желать лучшего. Для решения этой проблемы в 2015 г. В.С. Тимофеевым и А.А. Саниной была предложена новая модель, в основу которой легло обобщенно-нормальное семейство распределений [12]. Данное семейство при различных значениях параметров охватывает ряд симметричных распределений с тяжелыми и легкими хвостами, которые в некоторых случаях являются наиболее подходящими, чем нормальное или логистическое распределения. По причине того, что реальные данные зачастую невозможно описать симметричным законом, было принято решение о построении модели на основе другого универсального семейства распределений, а именно семейства устойчивых распределений, включающего в себя не только симметричные, но и несимметричные законы распределений, а также нормальное распределение при определенных значениях параметров. Цель работы – исследование новой модели с точки зрения качества классификации и выработка рекомендаций для применения усовершенствованной модели при решении практических задач.

1. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ И МЕТОДЫ РЕШЕНИЯ

Пусть проводится серия из m экспериментов. Обозначим за y_i некоторую зависимую переменную, принимающую одно из двух значений: 1 или 0 в зависимости от того, произошло или не произошло какое-то событие в конкретном эксперименте, $i = \overline{1,m}$. Полученные значения удобно объединить в вектор

$$Y = (y_1, \ldots, y_m),$$

где Y – вектор-столбец значений зависимой переменной. Выбор y_i может быть представлен в зависимости от значений входных факторов

$X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in R$ – значение j-го входного фактора для i-го наблюдения, $j = \overline{1,n}$, n – количество входных факторов. Для более удобного представления объединим значения входных факторов в проведенной серии экспериментов в матрицу

$$X = \left[x_{ij} \right].$$

Построим модель зависимости переменной Y от значений входных факторов с целью оценить вероятность наступления события y_i при различных значениях независимых переменных x_i. Основное уравнение модели

$$P(y_i = 1 \mid X_i) = F(z_i),$$

где $F(u)$ – кумулятивная функция одного из распределений, принимающая значения из отрезка $[0,1]$. Обычно в качестве $F(u)$ выбирается функция нормального или логистического распределений, значение которой изменяется в зависимости от величины z_i:

$$z_i = \theta_0 + \theta_1 x_{i1} + \ldots + \theta_n x_{in} = \theta(1,\, x_{i1},\, \ldots,\, x_{in})^T,$$

где z_i – линейная комбинация входных факторов; $\theta = (\theta_0, \ldots,\, \theta_n)$ – вектор-строка неизвестных коэффициентов (параметров модели).

Оценка параметров модели обычно производится методом максимального правдоподобия. Функция правдоподобия выражается как

$$L(\theta) = \prod_{i=1}^{m} F(z_i),$$

логарифмическая функция правдоподобия в этом случае равна

$$\ln(L(\theta)) = \sum_{i=1}^{m} \{ y_i \ln(F(z_i)) + (1 - y_i) \ln(1 - F(z_i)) \}.$$

Если в качестве функции $F(u)$ выбирается логистическое или нормальное распределение, получаются уже известные логит- и пробит-модели соответственно [3–9].

Рассмотрим решение простейшей задачи классификации с применением логит-модели. Пусть имеется два входных фактора x_1 и x_2 такие, что $x_1 \in [0,10]$, $x_2 \in [0,21]$. В качестве разделяющей гиперплоскости в данном случае будет выступать некоторая прямая $x_{2orig} = 2x_1 + 1$. Будем считать, что наблюдения, удовлетворяющие условию $x_2 > 2x_1 + 1$, относятся к первой группе, и обозначим их знаком ‘+’, как представлено на рис. 1. Пусть для этой группы наблюдений справедливо равенство $y_i = 0$. Наблюдения, удовлетворяющие неравенству $x_2 \le 2x_1 + 1$, попали во вторую группу и будут обозначены на рис. 1 знаком ‘**.**’. Для них будет верно $y_i = 1$. Таким образом были получены две незасоренные выборки с фиксированным набором значений входных факторов. Общее количество наблюдений в каждой такой выборке равно 25. Построение логит-модели для данного примера обеспечивает получение качественных результатов с нулевой ошибкой классификации.

Далее введем в рассматриваемую выборку два ошибочно классифицированных наблюдения, т. е. находящиеся по другую сторону от «своей» группы наблюдений относительно разделяющей группы гиперплоскости. Пример решения такой задачи с применением модели бинарного выбора, построенной на основе логистической функции, показан на рисунке, где x_{2logis} – уравнение, полученное в результате оценки параметров при построении модели логистической регрессии: $\theta = (-0.616,\ -0.145,\ 0.125)$. Как можно заметить, прямая x_{2logis} значительно отклонена от истинной прямой x_{2orig}, изначально разделяющей группы наблюдений.

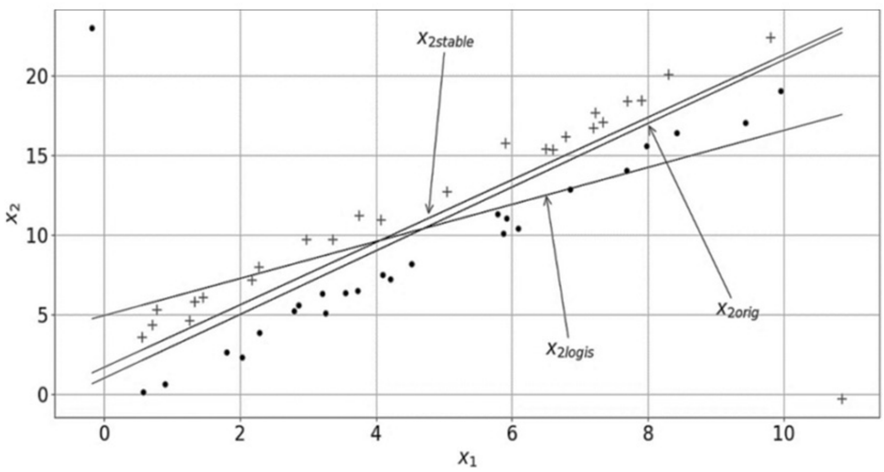
Формально качество полученного решения будем оценивать исходя из величины ошибки классификации

$$err = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{y}_i - y_i\right|,$$

где $\hat{y}_i$ – значение зависимой переменной, полученное в результате решения задачи классификации. Как указано в табл. 1, эта ошибка при решении рассмотренной задачи классификации с применением логит-модели составила около 30 %, что вряд ли следует считать удовлетворительным результатом.

Параметры уравнения $x_2 = kx_1 + c$ при решении задачи классификации с применением различных бинарных моделей выбора, а также величина ошибки классификации приведены в табл. 1.

Видно, что решение, полученное с применением логит-модели, оказалось очень неточным с точки зрения качества классификации в случае, когда среди исходных данных присутствует некоторое число ошибочно классифицированных наблюдений, расположенных за пределами эллипсоида рассеяния. Такие наблюдения в рамках данной статьи будем считать выбросами или аномальными наблюдениями.



Решение задачи классификации с применение различных функций для построения моделей бинарного выбора

Таблица 1

Параметры уравнения $x_2 = kx_1 + c$ при решении задачи классификации с применением бинарных моделей выбора

Модели выбора	k	c	err
x_{2orig}	2	1	0
x_{2logis}	1.160	4.928	0.28
$x_{2stable}$	1.964	1.676	0.04

В связи с тем, что логистическая регрессия показывает неудовлетворительный результат в смысле величины ошибки классификации, а выборки в общем случае могут быть засорены более чем двумя аномальными наблюдениями, было принято решение построить новую модель на основе семейства устойчивых распределений [13, 14]. При выборе $F(u)$ учитывались такие условия, как положительная определенность и непрерывность функции. Данные ограничения хорошо видны из логарифмической функции правдоподобия (1). Кроме того, хотелось иметь возможность представлять несимметричные законы распределений при изменении параметров всё той же функции. Указанным требованиям в значительной степени соответствует устойчивое семейство распределений, которое и было выбрано.

Устойчивое распределение полностью определяется характеристической функцией [14]

$$\varphi(t) = \exp\left[it\mu - |\sigma t|^{\alpha} (1 - i\beta \operatorname{sgn}(t)\Phi) \right],$$

$$\Phi = \begin{cases} \tan\left(\dfrac{\pi\alpha}{2}\right), \ \alpha \neq 1, \\ -\dfrac{2}{\pi}\lg|t|, \ \alpha = 1. \end{cases}$$

Здесь $t \in R$, $i = \sqrt{-1}$ – мнимая единица; α, β, σ, μ – параметры устойчивого распределения, удовлетворяющие условиям: $\alpha \in (0, 2]$ – параметр устойчивости, $\beta \in [-1, 1]$ – параметр асимметрии, $\sigma \in (0, +\infty)$ – параметр масштаба, $\mu \in (-\infty, +\infty)$ – параметр сдвига.

Далее для удобства объединим параметры распределения в вектор $p = (\alpha, \ \beta, \ \sigma, \ \mu)$.

Расчет кумулятивной функции распределения в этом случае происходит в два этапа.

1. Вычисление функции плотности распределения

$$f(x) = \frac{1}{2\pi} \int_{\Omega_R} \exp(-itx)\,\varphi(t)\,dt\,,$$

которая в дискретном виде представима как

$$f(x) = \frac{\Delta t}{2\pi} \sum_{k=1}^{K} \exp(-it_k x)\,\varphi(t_k)\,,$$

где Ω_R – область интегрирования функции плотности распределения, K – количество интервалов интегрирования, Δt – шаг дискретизации, $t_k = (k-1)\Delta t$, $\Delta t = \Omega_R / K$.

2. Вычисление функции распределения

$$F(u) = \int_{-\infty}^{u} f(x)\,dx\,.$$

Общий алгоритм решения задачи классификации можно представить в виде серии шагов:

• выбрать и зафиксировать некоторые начальные значения параметров устойчивого распределения;

• методом максимального правдоподобия найти неизвестные коэффициенты регрессии, решив задачу оптимизации для (1):

$$\hat{\theta} = \arg\max_{\theta}\, \ln(L(\theta))\,;$$

• вычислить ошибку классификации как среднее отклонение прогнозных значений от истинных значений

$$err = err\left(p, \hat{\theta}\right) = \frac{1}{m} \sum_{i=1}^{m} |\hat{y}_i - y_i|\,;$$

• решить дополнительную задачу оптимизации

$$\left(\hat{p}, \hat{\theta}\right) = \arg\min_{p}\, err\left(p, \hat{\theta}\right).$$

Дополнительная задача оптимизации путем варьирования значений параметров устойчивого распределения необязательна и здесь решается с целью улучшения качества классификации. Данный эффект достигается путем варьирования значений трех основных параметров устойчивого семейства распределения, которые, изменяя форму распределения, порождают частные законы распределений, наилучшим образом подстраивающиеся под входные данные.

При исследовании предложенной модели классификации было выявлено, что параметр регрессии θ_0 и параметр сдвига устойчивого распределения μ выполняют роль свободного коэффициента и для данной модели могут полностью заменять друг друга. Вследствие этого было решено зафиксировать параметр сдвига $\mu = 0$, тем самым снизив размерность пространства оптимизируемых параметров, с целью уменьшения временных затрат на работу алгоритма. Таким образом фактически оптимизация значений параметров выполнялась для вектора $(\alpha,\ \beta,\ \sigma,\ 0)$. В качестве вектора начальных приближений параметров устойчивого распределения в данной работе выбирается вектор $p = (1,\ 0.1,\ 1,\ 0)$, $\Omega_R \in [-\pi,\ \pi]$, $K = 2^6$.

Вернемся к рисунку и табл. 1, где $x_{2stable}$ – уравнение, полученное в результате оценки параметров при построении бинарной модели на основе устойчивого семейства распределения; err – величина ошибки классификации для решения задач с применением различных бинарных моделей. Видно, что модель, построенная на основе семейства устойчивого распределения, показывает значительно лучший результат классификации. Неизвестные параметры устойчивого распределения в этом случае получили оценки, приведенные в табл. 1, а вектор параметров устойчивого распределения $\hat{p} = (0.991,\ -0.538,\ 0.017,\ 0)$.

Далее проведем ряд вычислительных экспериментов на модельных данных с применением предложенного метода для решения задач классификации при различной степени засорения выборок аномальными наблюдениями. Перейдем к рассмотрению полученных результатов.

2. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Рассмотрим решение задачи классификации, аналогичной описанному выше примеру. Выбор y_i по-прежнему зависит от двух входных факторов. Данные $x_i = (x_{i1},\ x_{i2})$ зафиксированы, $x_1 \in [0,\ 10]$, $x_2 \in [0,\ 21]$, $i = \overline{1, m}$, $m = 500$,

$$y_i = \begin{cases} 1, \ F(u) > 0.5, \\ 0, \ F(u) \le 0.5, \end{cases}$$

Количество проводимых вычислительных экспериментов $N = 500$. Данные были засорены выбросами симметрично на 10, 20 и 50 процентов.

В табл. 2 представлены результаты решения задачи классификации при различной степени засорения выборки, при этом ошибочно классифицированные наблюдения находились по другую сторону от «своей» группы наблюдений относительно разделяющей гиперплоскости в зоне эллипсоида рассеяния. В табл. 3 представлены результаты решения аналогичной задачи классификации при различных уровнях зашумления данных с условием, что наблюдения-выбросы находились за пределами эллипсоида рассеяния.

В табл. 2 и 3 через F_Logit обозначена усредненная ошибка классификации $\overline{err}$ для серии экспериментов при решении задач с применением логит-модели; через F_Stable1 – усредненная ошибка классификации $\overline{err}$ для серии экспериментов при решении задач с применением модели, в основе которой лежит устойчивое распределение с некоторыми фиксированными значениями параметров; через F_Stable – усредненная ошибка классификации $\overline{err}$ для серии экспериментов при решении задач с применением модели, в основе которой лежит устойчивое распределение при варьировании параметров распределения.

Из табл. 2 видно, что в случае полного отсутствия или наличия малого количества аномальных наблюдений модели бинарного выбора, построенные на основе логистического и устойчивого распределений, оказываются эквивалентными в терминах ошибки классификации. И в том и в другом случае ошибка классификации равна нулю.

В случае засорения выборки некоторой долей выбросов более точные результаты решения задачи классификации показывает модель с устойчивым распределением в основе. Особенно это заметно при небольшой доле аномальных наблюдений (10…20 %, табл. 3).

Таблица 2

Ошибки классификации при различных уровнях засорения выборки аномальными наблюдениями, расположенными в зоне распределения группы

% засорения	F_Logit	F_Stable1	F_Stable2	F_Stable2-F_Logit
0	0.000	0.004	0.000	0.000
10	0.158	0.133	0.105	–0.053
20	0.237	0.288	0.249	0.012
50	0.200	0.208	0.187	–0.013

Таблица 3

Ошибки классификации при различных уровнях засорения выборки аномальными наблюдениями, расположенными за пределами зоны распределения группы

% засорения	F_Logit	F_Stable1	F_Stable2	F_Stable2-F_Logit
0	0.000	0.000	0.000	0.000
2	0.242	0.096	0.025	–0.217
4	0.312	0.125	0.080	–0.232
10	0.292	0.224	0.185	–0.107
20	0.248	0.239	0.213	–0.035
50	0.155	0.178	0.157	0.002

Таким образом, в ходе проведения серии экспериментов было установлено, что семейство устойчивых распределений имеет преимущество с точки зрения качества классификации при решении задач в случае, если данные неоднородны и имеют выбросы. Напомним, что под выбросом здесь понимаются наблюдения, расположенные за пределами области распределения группы, как показано на рисунке. Численно на небольшой группе экспериментов было показано, что при наличии в выборках такого рода выбросов бинарная модель выбора, построенная на основе семейства устойчивых распределений, при варьировании параметров последнего оказывается точнее с точки зрения величины ошибки классификации.

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрена задача классификации, решаемая с применением модели бинарного выбора. Традиционно для построения такой модели применяют нормальное или логистическое распределение. Однако ввиду того, что реальные данные представляют собой сложные структуры, которые плохо подчиняются упомянутым законам, порождая тем самым большую ошибку классификации, было принято решение о построении модели на основе более универсального семейства распределений, которое при варьировании параметров включает в себя множество симметричных и несимметричных законов, наилучшим образом описывающих данные.

Как показали результаты экспериментов, модель бинарного выбора, построенная на основе устойчивого семейства распределений при проведении экспериментов на выборочных данных, показывает преимущество в терминах минимизации ошибки классификации в решении задач данного класса. Особенно превосходство новой модели заметно в случае наличия в исходных данных небольшой доли выбросов – ошибочно классифицированных наблюдений.

Результат, полученный при решении задачи классификации с засорением исходных данных на 50 % при условии, что аномальные наблюдения расположены в области потенциально противоположной классификационной группы за пределами эллипсоида рассеяния, численно сравним для модели бинарного выбора построенной на основе логистической функции и устойчивого распределения. Отметим, что в таких случаях процент засорения слишком велик и сравним с объемом выборки, а учитывая характер выбросов, здесь должна решаться другая задача классификации.

Очевидно, что рассмотренный метод имеет существенное преимущество в сравнении с традиционными решениями (основанными на логит-моделях), особенно в случае засорения выборок аномальными наблюдениями указанного выше характера. Проведенные эксперименты подтверждают, что метод может использоваться на практике.

СПИСОК ЛИТЕРАТУРЫ

1. *Kropko J.* Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data: a thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Political Science. – Chapel Hill, 2008. – 46 p.

2. *Золотухин И.В.* Двухкомпонентное многомерное распределение Лапласа // Вестник Новгородского государственного университета им. Ярослава Мудрого. – 2012. – № 68. – С. 60–64.

3. *Judd Ch., McClelland G.H.* Data analysis: a model-comparison approach. – San Diego: Harcourt Brace Jovanovich, 1989. – 635 p.

4. *Форсайт Дж., Малькольм М.А., Моулер К.Б.* Машинные методы математических вычислений: пер. с англ. – М.: Мир, 1980. – 280 с.

5. *Каримов Р.Н.* Основы дискриминантного анализа: учебно-методическое пособие. – Саратов: Изд-во СГТУ, 2002. – 108 с.

6. *Rencher A.C.* Methods of multivariate analysis / Brigham Young University. – New York: John Wiley & Sons, 2002. – 727 p.

7. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.

8. *Кендалл М.Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды: пер. с англ. – М.: Наука, 1976. – 736 с.

9. *Каримов Р.Н.* Обработка экспериментальной информации. Ч. 3. Многомерный анализ: учебное пособие. – Саратов: Изд-во СГТУ, 2000. – 108 с.

10. *Press S.J., Wilson S.* Choosing between logistic regression and discriminant analysis // Journal of the American Statistical Association. – 1978. – Vol. 73, iss. 364. – P. 699–705. – doi: 10.1080/01621459.1978.10480080.

11. *Pohar M., Blas M., Turk S.* Comparison of logistic regression and linear discriminant analysis: a simulation study // Metodološki zvezki: advances in Methodology and Statistics. – 2004. – Vol. 1, N 1. – P. 143–161.

12. *Тимофеев В.С., Санина А.А.* Построение моделей бинарного выбора на основе универсального семейства распределений // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2015. – № 3. – С. 104–112.

13. *Chambers J.M., Mallows C.L., Stuck B.W.* A method for simulating stable random variables // Journal of the American Statistical Association. – 1976. – Vol. 71, iss. 354. – P. 340–344. – doi: 10.1080/01621459.1976.10480344.

14. *Золотарев В.М.* Одномерные устойчивые распределения. – М.: Наука, 1983. – 304 с.

Тимофеев Владимир Семенович, доктор технических наук, профессор кафедры теоретической и прикладной информатики факультета прикладной математики и информатики Новосибирского государственного технического университета. Основное направление научных исследований – разработка и исследование устойчивых методов и алгоритмов анализа многофакторных объектов, в том числе с использованием непараметрической статистики. Имеет более 100 публикаций, в том числе один учебник. E-mail: v.timofeev@corp.nstu.ru

Санина Анастасия Алексеевна, аспирант кафедры теоретической и прикладной информатики факультета прикладной математики и информатики Новосибирского государственного технического университета. Основное направление научных исследований – решение задач классификации: анализ существующих и разработка новых подходов. Имеет более 20 публикаций. E-mail: anastas.sanina@gmail.com

Binary choice modeling based on stable distribution[*]

V.S. TIMOFEEV[1], *A.A. SANINA*[2]

[1] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, D.Sc. (Eng), associate professor. E-mail: v.timofeev@corp.nstul.ru*
[2] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, post graduate student. E-mail: anastas.sanina@gmail.com*

The article describes the most popular issue which arises when solving classification problems. It is improving the classification model quality in terms of the number correctly classified objects. The most famous models (i.e. logit and probit models) which are usually used to solve classification problems have good results for limited data sets. So, the questions arise if it is possible to improve the classification quality and build a new model based on any other distribution which is more universal and convenient for real data instead of the logistic or normal distribution. An example is considered when the model includes two independent variables and is based on the logistic and then stable distribution. The results obtained are compared based on the classification error value. An improved approach to solve classification problems is proposed. This approach differs from the existing methods in two main aspects. The first one is that a new distribution is used to build a binary choice model. The second one is solving an additional optimization problem. The stable distribution parameters are changed in such a way as to minimize a classification error. Computational experiments are made to investigate the above approach. The results show that the method is effective. Classification quality is compared for different models with different contamination levels of the input data. The character of input data noisiness which greatly affects classification quality is revealed. Possible ways to further develop the above method are described. It should be noted that the proposed method is better than other approaches in terms of classification quality especially for data with a certain error type. Therefore the new method should be recommended for solving practical problems.

Keywords: classification problem, logit-model, binary choice model, stable distribution, classification error, likelihood function, outline, noise data, characteristic function, computational experiment, parameter estimation, distribution function

DOI: 10.17212/1814-1196-2017-4-105-116

REFERENCES

1. Kropko J. *Choosing between multinomial logit and multinomial probit models for analysis of unordered choice data:* a thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Political Science. Chapel Hill, 2008. 46 p.

2. Zolotukhin I.V. Dvukhkomponentnoe mnogomernoe raspredelenie Laplasa [Two-component multivariate Laplace distribution]. *Vestnik Novgorodskogo gosudarstvennogo universiteta – Vestnik of Yaroslav the wise Novgorod state university*, 2012, no. 68, pp. 60–64.

3. Judd Ch., McClelland G.H. *Data analysis: a model-comparison approach*. San Diego, Harcourt Brace Jovanovich, 1989. 635 p.

4. Forsythe G.E., Malcolm M.A., Moler C.B. *Computer methods for mathematical computations*. New Jersey, Prentice-Hall, 1977. 270 p. (Russ. ed.: Forsait Dzh., Mal'kol'm M., Mouler K. *Mashinnye metody matematicheskikh vychislenii*. Translated from English. Moscow, Mir Publ., 1980. 280 p.).

5. Karimov R.N. *Osnovy diskriminantnogo analiza* [Fundamentals of discrimi-nant analysis]. Saratov, SSTU Publ., 2002. 108 p.

6. Rencher A.C. *Methods of multivariate analysis*. Brigham Young University. New York, John Wiley & Sons, 2002. 727 p.

[*] *Received 27 October 2017.*

7. Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: klassifikatsiya i snizhenie razmernosti* [Applied statistics: classification and size decrease]. Moscow, Finansy I statistika Publ., 1989. 607 p.

8. Kendall M.G., Stuart A. *The advanced theory of statistics.* Vol. 3. *Design and analysis and thime series.* 2nd ed. London, Charles Griffin and Company, 1968 (Russ. ed.: Kendall M.Dzh., St'yuart A. *Mnogomernyi statisticheskii analiz i vremennye ryady*. Moscow, Nauka Publ., 1976. 736 p.).

9. Karimov R.N. *Obrabotka eksperimental'noi informatsii.* Ch. 3. *Mnogomernyi analiz* [Processing of experimental data. Pt. 3. Multivariate analysis]. Saratov, SSTU Publ., 2000. 108 p.

10. Press S.J., Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 1978, vol. 73, iss. 364, pp. 699–705. doi: 10.1080/01621459.1978.10480080.

11. Pohar M., Blas M., Turk S. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodološki zvezki: Advances in Methodology and Statistics*, 2004, vol. 1, no. 1, pp. 143–161.

12. Timofeev V.S., Sanina A.A. Postroenie modelei binarnogo vybora na osnove universal'nogo semeistva raspredelenii [Binary choice modelling based on the universal distribution]. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Upravlenie, vychislitel'naya tekhnika i informatika – Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*, 2015, no. 3, pp. 104–112.

13. Chambers J.M., Mallows C.L., Stuck B.W. A method for simulating stable random variables. *Journal of the American Statistical Association*, 1976, vol. 71, iss. 354, pp. 340–344. doi: 10.1080/01621459.1976.10480344.

14. Zolotarev V.M. *Odnomernye ustoichivye respredeleniya* [One-dimensional stable distribution]. Moscow, Nauka Publ., 1983, 304 p.

Для цитирования:

Тимофеев В.С., Санина А.А. Построение модели бинарного выбора на основе устойчивого семейства распределений // Научный вестник НГТУ. – 2017. – № 4 (69). – С. 105–116. – doi: 10.17212/1814-1196-2017-4-105-116.

For citation:

Timofeev V.S., Sanina A.A. Postroenie modeli binarnogo vybora na osnove ustoichivogo semeistva raspredelenii [Binary choice modeling based on stable distribution]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2017, no. 4 (69), pp. 105–116. doi: 10.17212/1814-1196-2017-4-105-116.