УДК 519.816

В. С. Тимофеев, А. А. Санина

ПОСТРОЕНИЕ МОДЕЛЕЙ БИНАРНОГО ВЫБОРА НА ОСНОВЕ УНИВЕРСАЛЬНОГО СЕМЕЙСТВА РАСПРЕДЕЛЕНИЙ[1]

Рассмотрена задача классификации и некоторые распространённые методы её решения с применением моделей бинарного выбора. Среди предложенных моделей предпочтение отдаётся logit- и probit-моделям в связи с возможностью обрабатывать входные факторы различной природы. Рассматривается закономерный вопрос о возможности введения новой модели, в основе которой будет лежать универсальное семейство распределений, отличное от логистического закона для logit-модели и нормального закона – для probit-модели. Подробно описывается математическая постановка, приводятся пояснения, касающиеся возможности введения новой модели, обозначены существующие требования и ограничения. Кроме того, предложен новый метод оценивания параметров классифицирующей функции, основанный на применении универсального распределения. В качестве такого распределения предлагается использовать обобщённое нормальное распределение. Предложенная процедура классификации заключается в решении двойной задачи оптимизации: минимизации функции правдоподобия при подборе оптимальных коэффициентов для классифицирующей функции и минимизации значения величины ошибки классификации путём варьирования параметров выбранного распределения. С целью исследования предложенного метода проведён ряд вычислительных экспериментов при различных объёмах выборок, количестве входных факторов и различных зависимостях в исходных данных. Результаты экспериментов подробно изучены с целью выявить влияние распределения входных переменных на характер эмпирического распределения вероятностной модели. Полученные результаты свидетельствуют об эффективности предложенной процедуры. Особенно хорошо это иллюстрируют тесты на расширенной модели (с большим количеством переменных). Указаны возможные перспективы развития работы: в связи с тем, что предложенный метод прошёл ряд испытаний, в дальнейшем можно исследовать величину ошибки классификации, выбирая для построения модели любые другие распределения при соблюдении некоторых условий. Немаловажно, что усовершенствованный метод решения задач классификации даёт значительное улучшение качества классификации существующих процедур, а соответственно, может быть рекомендован для применения на практике.

Ключевые слова: дискриминантный анализ, logit-модель, probit-модель, функция правдоподобия, задача классификации, факторы, бинарная зависимая переменная, процедура оптимизации, обобщённое нормальное распределение.

Введение

Классификация наблюдений (или задача принятия решения) находит своё применение при решении задач в профессиональной деятельности. Класс таких задач представляет собой множество постановок, когда исследователь имеет некоторый объект (это может быть товар, услуга и др.) и набор характеризующих его признаков (например, цена, качество и др.), взвесив которые он принимает некоторое решение относительно имеющегося объекта (купить, продать и др.). В настоящее время для решения задач подобного класса применяются математические модели дискретного выбора: logit- и probit-модели, частным случаем которых выступает модель дискриминантного анализа [1–11]. Требование по выполнению основных предположений дискриминантного анализа, таких как непрерывность, независимость и нормальное распределение входных факторов в реальных условиях, сильно ограничивает круг задач, для решения которых может применяться эта модель [12, 13]. Logit- и probit-модели менее требовательны к входным данным, а следовательно, являются более гибкими [12, 13].

Для того чтобы сделать выбор в пользу какой-либо модели, необходимо исследовать каждую из них в отдельности на заданных наборах данных с точки зрения качества классификации. Кроме того, зная, что в основе logit- и probit-моделей лежат логистическое и нормальное распределения соответственно, разумно предположить, что можно построить модель с каким-либо другим распределением в её основе. Таких распределений существует великое множество, и в идеале хотелось бы иметь качественную универсальную модель для решения широкого спектра задач. Для достижения желаемого результата предлагается строить новую модель на основе некоторого

[1] Работа выполнена при финансовой поддержке Министерства образования и науки РФ по государственному заданию № 2014/138, проект № 1689.

универсального семейства распределений. Следует отметить, что ранее универсальные семейства распределений уже использовались при построении регрессионных зависимостей (см., например, [14, 15]). Это решение разумно ввиду того, что универсальные семейства имеют большое варьирование форм, среди которых существуют и хорошо известные законы распределений. В качестве такого «оптимального» закона мы, в рамках нашего исследования, предлагаем выбрать семейство обобщённого нормального распределения. При варьировании собственных параметров это распределение описывает такие известные частные законы, как нормальный и логистический. Исследование новой модели проводится с точки зрения качества классификации и сравнения полученных результатов с работой logit-модели.

Постановка задачи

Пусть зависимая переменная y принимает одно из двух значений: 0 или 1 в зависимости от наступления или ненаступления некоторого события. В качестве вектора значений входных признаков для каждого i-го наблюдения будем рассматривать вектор $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $x_{ij} \in R$ – значение j-го фактора для i-го наблюдения, $i = \overline{1,m}$, $j = \overline{1,n}$. Для описания зависимости вероятности наступления события $y = 1$ от входных факторов построим модель, основное уравнение которой записывается в виде

$$P\{y_i = 1 \mid x_i\} = F(z_i),$$

где $F(u)$ – некоторая функция. Обычно в качестве $F(u)$ используют одну из функций распределения, величина z_i определяется как линейная комбинация входных факторов:

$$z_i = \theta x_i^T = \theta_1 x_{i1} + \ldots + \theta_n x_{in},$$

где $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ – вектор неизвестных параметров.

Методы решения

Оценивание $\theta_1, \theta_2, \ldots, \theta_n$ проводится по набору значений независимых переменных и соответствующих им значений зависимой переменной y. Обычно для этого используется метод максимального правдоподобия, согласно которому необходимо максимизировать значение функции правдоподобия. Однако на практике удобнее использовать логарифмированное выражение для функции правдоподобия:

$$\ln L(\theta) = \sum_{i=1}^{m} y_i \ln F\left(\theta x_i^T\right) + (1 - y_i) \ln\left(1 - F\left(\theta x_i^T\right)\right). \qquad (1)$$

Традиционно в качестве $F(u)$ выбирается логистическая или нормальная функция распределения, в этом случае мы получим logit- и probit-модель соответственно. Однако, если вероятность наступления некоторого события описывается законом, отличным от логистического и нормального, качество классификации будет ухудшаться в зависимости от характера различия эмпирического и модельного законов распределения. Именно поэтому предлагается строить модель, основанную на семействе одного из универсальных распределений. Преимуществом этого выбора является то, что такие семейства распределений при варьировании собственных параметров имеют частными случаями некоторые уже известные законы распределения. Из формулы (1) видно, что в качестве $F(u)$ необходимо выбирать распределения, определённые на всей действительной оси, т. е. $z \in R$. В связи с этим было принято решение использовать обобщённое нормальное распределение с неизвестными параметрами μ, $\alpha \in R$, $\beta \geq 0$:

$$f(z, \mu, \alpha, \beta) = \frac{1}{2\alpha\Gamma(1 + 1/\beta)} \exp\left[-\left(\frac{|z - \mu|}{\alpha}\right)^{\beta}\right].$$

Обобщённое нормальное распределение представляет собой параметрическое семейство распределений. Оно включает в себя нормальное распределение, распределение Лапласа, а также равномерное распределение на ограниченных интервалах действительной прямой. Распреде-

ление из данного семейства является нормальным при $\beta = 2$ (с математическим ожиданием μ и дисперсией $\frac{\alpha^2}{2}$) и является распределением Лапласа при $\beta = 1$. Данное семейство демонстрирует наличие хвостов распределения, которые тяжелее нормальных при $\beta < 2$ и легче нормальных при $\beta > 2$ [16].

Для оценки качества классификации была использована доля неверно классифицированных наблюдений, которая, с учетом специфики постановки, может быть вычислена следующим образом:

$$Err = Err\left(\hat{\theta}, \mu, \alpha, \beta\right) = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{y}_i - y_i\right|,$$

где $\hat{y}_i$ – прогнозное значение зависимой переменной. Так как обобщенное нормальное распределение зависит от параметров, то вместо $F(u)$ следует использовать $F\left(z, \mu, \alpha, \beta\right)$. Выбирая значения неизвестных параметров специальным образом, можно проводить процедуру классификации на качественно более высоком уровне. Очевидно, что при варьировании значений параметров будет изменяться значение величины Err. Таким образом, возникает ещё одна задача: минимизировать данную ошибку классификации:

$$\left(\mu, \alpha, \hat{\beta}\right) = \underset{(\mu, \alpha, \beta)}{\arg\min}\left(Err\left(\hat{\theta}, \mu, \alpha, \beta\right)\right). \qquad (2)$$

Следует отметить: возможны случаи, когда рассмотренные модели не будут работать совсем в связи с тем, что при определённых значениях факторов и коэффициентов (значения параметров $\theta_1, \theta_2, ..., \theta_n$) аргумент функции $F(u)$ может оказаться «слишком большим» или наоборот. Функция примет свои экстремальные значения, которые несовместимы с корректной обработкой функции правдоподобия при подборе неизвестных параметров методом максимального правдоподобия. В этом случае рекомендуется провести предварительную нормировку входных факторов.

Далее рассмотрим точность классификации новой модели и традиционной модели логистической регрессии.

Результаты экспериментов

Исследование работоспособности logit-модели и модели, построенной на основе обобщённого нормального распределения с дополнительной процедурой оптимизации (2), проводилось на основе вычислительных экспериментов. С целью исследования качества классификации при различных зависимостях в исходных данных независимые переменные представлялись как выборки из следующих непрерывных законов распределения: нормальный – N, экспоненциальный – Exp, обобщённое нормальное распределение с тяжёлыми и лёгкими хвостами (GN(1), GN(10)). Выходная переменная – бинарная случайная величина с вероятностью успеха, моделируемой на основе нормального закона распределения. Количество наблюдений, соответствующих значению $y = 0$, равно m_1, а наблюдений, соответствующих значению $y = 1$, – соответственно m_2. В рамках нашего исследования полагалось $m_1 = m_2$. Общее количество наблюдений $m =$ 50, 100, 200, 500, при этом очевидно, что $m = m_1 + m_2$.

В табл. 1–3 приведены значения показателя *Err* при решении задач классификации по оценённым значениям параметров уравнения (1) и параметров обобщённого нормального закона распределения. Обозначения, принятые в таблицах: F_Logit – при построении модели использована logit-функция; F_GN1 – при построении модели использована функция обобщённого нормального закона распределения при фиксированных значениях параметров ($\mu = 0$, $\alpha = 1$, $\beta = 0$); F_GN2 – при построении модели использована функция обобщённого нормального закона распределения с решением дополнительной задачи оптимизации (2). Probit-модель была исключена из рассмотрения в связи с тем, что на более ранних этапах исследования при решении задач классификации с её применением был получен результат, эквивалентный (или ещё менее точный) результату, полученному при решении задач с применением logit-модели.

Таблица 1

Значения показателя *Err* для модели с одной переменной

Закон	*m*	F_Logit	F_GN1	F_GN2	F_GN2 –F_Logit	F_Logit/ F_GN2
N	50	0.00E+00	0.00E+00	0.00E+00	0.00E+00	–
	100	0.00E+00	0.00E+00	0.00E+00	0.00E+00	–
	200	0.00E+00	0.00E+00	0.00E+00	0.00E+00	–
	500	0.00E+00	4.00E-06	0.00E+00	0.00E+00	–
Exp	50	0.00E+00	4.00E-05	4.00E-05	4.00E-05	0.000
	100	2.00E-05	4.00E-05	0.00E+00	-2.00E-05	–
	200	2.00E-05	2.00E-05	2.00E-05	0.00E+00	1.000
	500	4.00E-05	6.80E-05	4.00E-06	-3.60E-05	10.000
GN (1)	50	0.00E+00	2.00E-04	4.00E-05	4.00E-05	0.000
	100	1.00E-04	1.80E-04	2.00E-05	-8.00E-05	5.000
	200	9.00E-05	2.60E-04	3.00E-05	-6.00E-05	3.000
	500	1.88E-04	3.16E-04	2.00E-05	-1.68E-04	9.400
GN (10)	50	8.00E-04	9.20E-04	0.00E+00	-8.00E-04	–
	100	7.20E-04	7.80E-04	0.00E+00	-7.20E-04	–
	200	8.80E-04	8.80E-04	8.80E-04	0.00E+00	1.000
	500	9.68E-04	9.76E-04	3.60E-05	-9.32E-04	26.889

Таблица 2

Значения показателя *Err* для модели с тремя переменными

Закон	*m*	F_Logit	F_GN1	F_GN2	F_GN2 –F_Logit	F_Logit/ F_GN2
N	50	4.00E-05	4.00E-05	8.00E-05	4.00E-05	0.500
	100	0.00E+00	4.00E-05	0.00E+00	0.00E+00	–
	200	0.00E+00	0.00E+00	5.00E-05	5.00E-05	0.000
	500	0.00E+00	6.40E-05	4.00E-06	4.00E-06	0.000
Exp	50	0.00E+00	0.00E+00	0.00E+00	0.00E+00	–
	100	2.00E-05	4.00E-05	2.00E-05	0.00E+00	1.000
	200	3.00E-05	1.40E-04	4.00E-05	1.00E-05	0.750
	500	2.40E-05	2.80E-05	1.20E-05	-1.20E-05	2.000
GN (1)	50	0.00E+00	1.60E-04	8.00E-05	8.00E-05	0.000
	100	6.00E-05	3.40E-04	8.00E-05	2.00E-05	0.750
	200	4.00E-05	1.80E-04	1.00E-05	-3.00E-05	4.000
	500	1.64E-04	3.04E-04	4.00E-06	-1.60E-04	41.000
GN (10)	50	4.80E-04	6.40E-04	1.20E-04	-3.60E-04	4.000
	100	6.00E-04	7.20E-04	2.00E-04	-4.00E-04	3.000
	200	6.70E-04	8.20E-04	1.00E-05	-6.60E-04	67.000
	500	8.88E-04	9.20E-04	1.20E-04	-7.68E-04	7.400

Таблица 3

Значения показателя *Err* для модели с пятью переменными

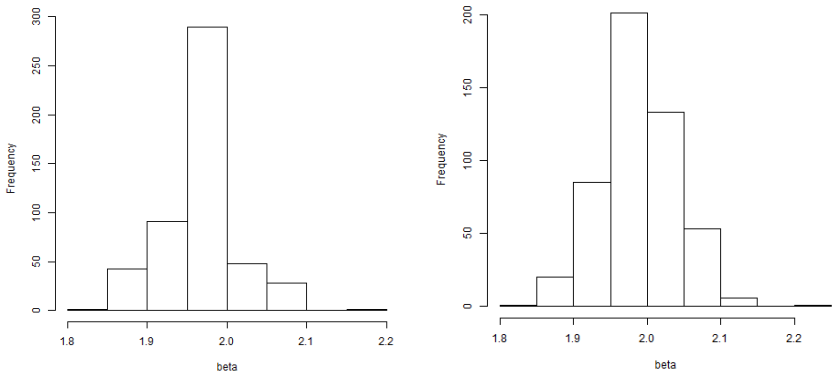
Закон	*m*	F_Logit	F_GN1	F_GN2	F_GN2 –F_Logit	F_Logit/ F_GN2
N	50	4.00E-05	8.00E-05	0.00E+00	-4.00E-05	–
	100	0.00E+00	6.00E-05	0.00E+00	0.00E+00	–
	200	6.00E-05	1.30E-04	0.00E+00	-6.00E-05	–
	500	1.52E-04	5.20E-05	1.60E-05	-1.36E-04	9.500
Exp	50	1.20E-04	8.00E-05	8.00E-05	-4.00E-05	1.500
	100	2.00E-04	1.00E-04	2.00E-05	-1.80E-04	10.000
	200	8.00E-05	7.00E-05	3.00E-05	-5.00E-05	2.667
	500	1.00E-04	2.24E-04	6.40E-05	-3.60E-05	1.563
GN (1)	50	2.00E-04	2.00E-04	2.00E-04	0.00E+00	1.000
	100	1.40E-04	2.80E-04	1.40E-04	0.00E+00	1.000
	200	3.70E-04	3.30E-04	1.50E-04	-2.20E-04	2.467
	500	3.44E-04	3.40E-04	2.40E-04	-1.04E-04	1.433
GN (10)	50	6.80E-04	7.20E-04	6.00E-04	-8.00E-05	1.133
	100	8.20E-04	8.40E-04	5.00E-04	-3.20E-04	1.640
	200	8.50E-04	9.40E-04	6.60E-04	-1.90E-04	1.288
	500	8.88E-04	8.88E-04	8.84E-04	-4.00E-06	1.005

Из приведённых выше таблиц видно, что с точки зрения качества классификации при больших объёмах выборки (m = 500) решение задачи (2) оказывается стабильно лучше решения с применением logit-модели (в среднем в 9,2 раза). Исключение составляет случай, когда вход-

ные факторы распределены по нормальному закону (количество факторов менее пяти) – результат решения задачи (2) эквивалентен решению задачи классификации с применением logit-модели. При увеличении количества переменных с одной до трёх и пяти, модель, построенная на основе обобщённого нормального закона распределения, показывает лучшее решение на большем наборе тестов. Дополнительная процедура подбора параметров семейства распределения позволила улучшить решение задачи классификации с применением logit-модели до 10 раз при расширенном наборе факторов (количество факторов – 5).

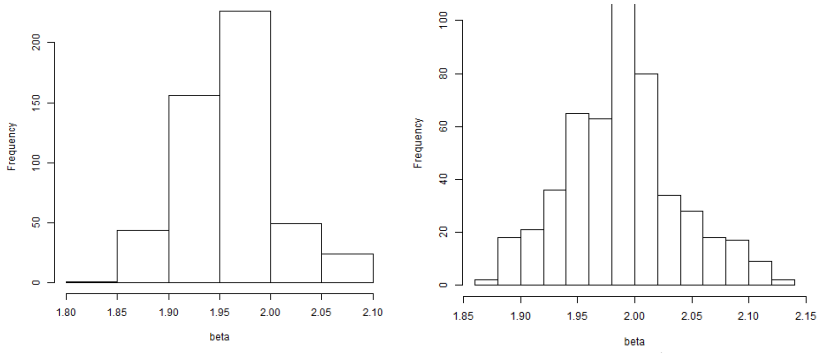
В целом усовершенствованный метод классификации показывает лучшее решение в сравнении со стандартной logit-моделью на большинстве рассмотренных наборов данных. В среднем это лучше в 4,7 раза.

Как уже было сказано выше, значение параметра формы β обобщенного нормального закона распределения позволяет получить информацию о характере отклонения итогового распределения от нормального в сторону тяжёлых или лёгких хвостов. Рассмотрим подробнее вид распределений оценённых значений параметра формы, полученных в ходе вычислительных экспериментов при различных условиях. В случае, когда признаки распределены по нормальному закону при больших объёмах выборок ($m \geq 200$) и количестве переменных менее пяти, хвосты закона распределения, наилучшим образом описывающего вероятностную модель, становятся тяжёлыми. Это явление продемонстрировано на рис. 1, 2. На горизонтальной и вертикальной осях отображены сгруппированные значения параметра β и частота их появления соответственно. Чем выше столбцы гистограммы, расположенные левее $\beta = 2$, тем тяжелее хвосты итогового распределения, и наоборот, чем больше количество столбцов и их высота на промежутке $\beta > 2$, тем легче хвосты эмпирического распределения, наилучшим образом описывающего вероятностную модель.



а *б*

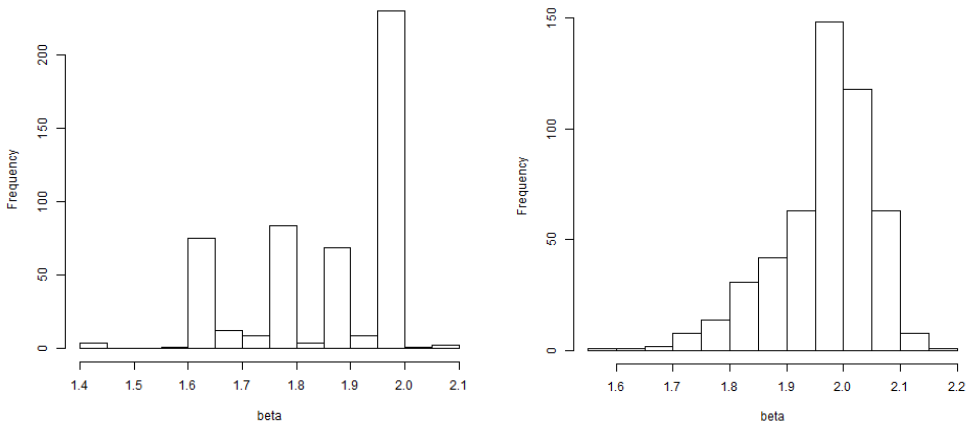
Рис. 1. Параметр β для моделей:
а – с тремя; *б* – с пятью переменными (m = 200)



а *б*

Рис. 2. Параметр β для моделей:
а – с тремя; *б* – с пятью переменными (m = 500)

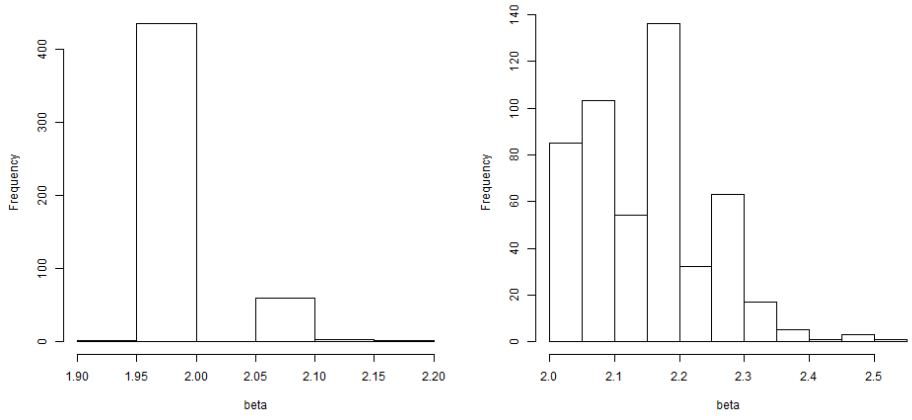
Если входные факторы – выборка из закона распределения с тяжёлыми хвостами, итоговая модель содержит распределения с тяжёлыми хвостами, независимо от количества факторов и объёмов выборок (рис. 3, *а*, *б*).



а *б*

Рис. 3. Параметр β для моделей:
а – с одной переменной ($m = 500$); *б* – с пятью переменными ($m = 500$)

При распределении независимых переменных по закону с лёгкими хвостами, характер закона распределения в основе вероятностной модели сохраняется аналогично случаю, описанному ранее: наилучшую модель образуют распределения с лёгкими хвостами. Исключение составляют выборки больших объёмов ($m = 500$) для расширенной модели с пятью факторами, когда отклонение от нормального закона распределения, образующего модель, симметрично в сторону лёгких и тяжёлых хвостов. Такой же симметричный характер отклонения наблюдается и в случае распределения независимых переменных согласно несимметричному закону. Исключением является лишь модель с одной переменной, когда отклонение от нормального закона для распределения, образующего модель, несимметрично и смещено в сторону распределения с лёгкими хвостами (рис. 4, *а*, *б*).



а *б*

Рис. 4. Параметр β для модели с одной переменной:
а – $m = 50$; *б* – $m = 500$

Подводя итог, можно сделать следующий вывод: особенности распределения входных факторов приводят к тому, что наилучшее распределение, описывающее эмпирические данные,

получается при различных значениях параметра формы β, а это далеко не всегда соответствует logit- и probit-моделям. Таким образом, использование для построения вероятностной модели обобщённого нормального семейства распределений повышает качество классификации.

Заключение

Таким образом, в работе предложена новая модель дискретного выбора, построенная на основе универсального семейства распределений. В качестве такого семейства было выбрано обобщённое нормальное распределение. Данная модель является обобщением существующих частных моделей бинарного выбора: logit- и probit-моделей. Соответственно, в дальнейшем возможен поиск других универсальных семейств распределения для построения обобщённых моделей и исследование их с точки зрения качества классификации для решения рассмотренной задачи.

При варьировании параметра формы β семейство обобщённого нормального распределения описывает частные случаи других законов распределения с лёгкими и тяжёлыми хвостами, обеспечивая тем самым более точный результат при решении задачи классификации. Учитывая всё сказанное ранее, новую модель и усовершенствованный алгоритм классификации можно рекомендовать для применения 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 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 / J. Kropko. Chapel Hill, 2008. 46 p.

2. *Золотухин И. В.* Двухкомпонентное многомерное распределение Лапласа / И. В. Золотухин // Вестн. Новгород. гос. ун-та им. Ярослава Мудрого. 2012. № 68. С. 60–64.

3. *Малхотра Н. К.* Маркетинговые исследования: практическое руководство / Н. К. Малхотра. М.: Изд. дом «Вильямс», 2002. 960 с. + Прил. (1 CD-ROM).

4. *StatSoft.* Электронный учебник по статистике. М., 2012 // URL: http://www.statsoft.ru/home/textbook/default.htm. 2005 (дата обращения: 02.02.2015).

5. *Judd Ch.* Data analysis / Ch. Judd, G. McCleland. Harcourt Brace Jovanovich, USA, 1989. 107 p.

6. *Форсайт Дж.* Машинные методы математических вычислений / Дж. Форсайт, М. Малькольм, К. Моулер. М.: Мир, 1980. 280 с.

7. *Каримов Р. Н.* Основы дискриминантного анализа: учеб.-метод. пособие / Р. Н. Каримов. Саратов: Изд-во СГТУ, 2002. 108 с.

8. *Rencher A. C.* Methods of multivariate analysis / A. C. Rencher. Brigham Young University, USA, 2002. 727 p.

9. *Айвазян С. А.* 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.

10. *Кендалл М. Дж.* Многомерный статистический анализ и временные ряды / М. Дж. Кендалл, А. М. Стьюарт. М.: Наука, 1976. 736 с.

11. *Каримов Р. Н.* Обработка экспериментальной информации: учеб. пособие. Ч. 3. Многомерный анализ // Р. Н. Каримов. Саратов: Изд-во СГТУ, 2000. 108 с.

12. *Press S. J.* Choosing between logistic regression and discriminant analysis / S. J. Press, S. Wilson // Journal of the America Statistical Assotiation. 1978. Vol. 73, iss. 364. P. 699–705.

13. *Pohar M.* Comparison of logistic regression and linear discriminant analysis: a simulation study / M. Pohar, M. Blas, S. Turk // Metodolovski zvezki journal: advances in Methodilogy and Statistics. 2004. Vol. 1, no. 1. P. 143–161.

14. *Тимофеев В. С.* Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда-распределения / В. С. Тимофеев, Е. А. Хайленко // Докл. Акад. наук высш. шк. РФ. Новосибирск: Изд-во НГТУ, 2010. № 2 (15). С. 25–36.

15. *Тимофеев В. С.* Оценивание параметров регрессионных зависимостей с использованием кривых Пирсона / В. С. Тимофеев // Науч. вестн. Новосибир. гос. техн. ун-та. 2009. № 4 (37). С. 57–66.

16. *Денисов В. И.* Методы построения многофакторных моделей по неоднородным, негауссовским, зависимым наблюдениям / В. И. Денисов, Д. В. Лисицин. Новосибирск: Изд-во НГТУ, 2008. 360 с.

Статья поступила в редакцию 9.06.2015

ИНФОРМАЦИЯ ОБ АВТОРАХ

Тимофеев Владимир Семёнович – Россия, 630073, Новосибирск; Новосибирский государственный технический университет; д-р техн. наук, доцент; профессор кафедры «Теоретическая и прикладная информатика»; v.timofeev@corp.nstu.ru.

Санина Анастасия Алексеевна – Россия, 630073, Новосибирск; Новосибирский государственный технический университет; аспирант кафедры «Теоретическая и прикладная информатика»; anastas.sanina@gmail.com.

V. S. Timofeev, A. A. Sanina

BINARY CHOICE MODELLING BASED ON THE UNIVERSAL DISTRIBUTION

Abstract. The paper considers the problem of classification and some methods for its solution based on the binary choice models. Logit- and probit models have been preferred to discriminant function model because they are able to process different input data types. So, the question on the possible introduction of the new model based on the function, which differs from the logit function for the logit model and the normal function for probit model respectively, is considered. The mathematical model is fully described, the possibility of introduction of a new model is justified and the existing restrictions preventing this action are given. Moreover, a new method for evaluation of the parameters of the classification function, based on the universal distribution, is presented. It is proposed to take the general normal distribution as a new distribution with unknown parameters. The new classification procedure helps solve the dual optimization problem: minimization of the likelihood function with the optimal coefficients fitting for the classification function and minimization of the classification error magnitude by varying the parameters of the selected distribution. In order to test the new method, a set of computational experiments was performed with different sample sizes and varied number of income variables and various dependencies in the input data. The results were studied in detail in order to fix the influence of input data distribution on the probability model empirical distribution. The obtained results show the effectiveness of the proposed procedure. This is particularly well observed in the tests with the extended model (with a lot of variables). The possible ways of further development of the work are noted. Due to the fact, that the proposed method works well, it is possible to study the magnitude of the classification error by choosing any other statistical distribution for creating the models with the certain conditions in the future. It should be noted, that the new method for solving the classification problem significantly improves the classification quality of the existing procedures, so it can be successfully applied in practice.

Key words: discriminant analysis, logit model, probit model, likelihood function, classification problem, factors, two-valued dependent variable, optimization procedure, general normal distribution.

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 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 imeni Iaroslava Mudrogo*, 2012, no. 68, pp. 60–64.

3. Malhotra N. K. *Marketing research: an applied approach.* Harlow, England, London, New York, Financial Times, Prentice Hall, 2002. 816 p. Includes CD-ROM (Russ. Ed.: Malhotra N. K. *Marketingovye issledovaniya: prakticheskoe rukovodstvo.* Translated from English. Moscow, Williams Publ., 2002. 957 p. + Prilozhenie 1 CD-ROM).

4. *StatSoft. Elektronnyi uchebnik po statistike* [Electronic textbook on Statistics]. StatSoft. Moscow, 2012 // URL: http://www.statsoft.ru/home/textbook/default.htm. 2005 (accessed: 02.02.2015).

5. Judd Ch., McCleland G. *Data analysis*. Harcourt Brace Jovanovich, USA, 1989. 107 p.

6. Forsythe G. E., Malkolm M. A., Mouler 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*]. Moscow, Mir Publ., 1980. 280 p.

7. Karimov R. N. *Osnovy diskriminantnogo analiza* [Fundamentals of discriminant analysis]. Saratov, Izd-vo SGTU, 2002. 108 p.

8. Rencher A. C. *Methods of multivariate analysis*. Brigham Young University, USA, 2002. 727 p.

9. Aivazian S. A., Bukhshtaber V. M., Eniukov I. S., Meshalkin L. D. *Prikladnaia statistika: klassifikatsiia i snizhenie razmernosti* [Applied statistics: classification and size decrease]. Moscow, Finansy i statistika Publ., 1989. 607 p.

10. Kendall M. G., Stuart A. *The advanced theory of statistics*. Vol. 3. Design and Analysis and time series. London, Charles Griffin and Company, 1968. 736 p. (Russ. ed.: Kendall M. Dzh., St'iuart A. M. Mnogomernyi statisticheskii analiz i vremennye riady. Moscow, Nauka Publ., 1976. 736 p.

11. Karimov R. N. *Obrabotka eksperimental'noi informatsii* [Processing of experimental data. Pt. 3. Multivariate analysis]. Saratov, Izd-vo SGTU, 2000. 108 p.

12. Press S. J., Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the America Statistical Assotiation*, 1978, vol. 73, iss. 364, pp. 699–705.

13. Pohar M., Blas M., Turk S. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodolovski zvezki journal: advances in Methodilogy and Statistics*, 2004, vol. 1, no. 1, pp. 143–161

14. Timofeev V. S., Khailenko E. A. Adaptivnoe otsenivanie parametrov regressionnykh modelei s ispol'zovaniem obobshchennogo liambda-raspredeleniia [Adaptive estimation of regression model parameters with error distribution inhomogeneity]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii*, 2010, no. 2 (15), pp. 25–36.

15. Timofeev V. S. Otsenivanie parametrov regressionnykh zavisimostei s ispol'zovaniem krivykh Pirsona [The Pirson's curves in parameter estimation problem for regression model]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta*, 2009, no. 4 (37), pp. 57–66.

16. Denisov V. I., Lisitsin D. V. *Metody postroeniia mnogofaktornykh modelei po neodnorodnym, negaussovskim, zavisimym nabliudeniiam* [Constructing Methods for the Multiple Models Based on the Inhomogeneous Non-Gaussian Dependent Data]. Novosibirsk, Izd-vo NGTU, 2008. 360 p.

The article submitted to the editors 9.06.2015

INFORMATION ABOUT THE AUTHORS

Timofeev Vladimir Semenovich – Russia, 630073, Novosibirsk; Novosibirsk State Technical University; Doctor of Technical Sciences, Assistant Professor; Professor of the Department "Theoretical and Applied Computer Science"; v.timofeev@corp.nstu.ru.

Sanina Anastasiia Alekseevna – Russia, 630073, Novosibirsk; Novosibirsk State Technical University; Postgraduate Student of the Department "Theoretical and Applied Computer Science", anastas.sanina@gmail.com.