СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2015. – № 1(79).** – 23–32

УДК 519.816

ПОВЫШЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ЛИНЕЙНЫХ МОДЕЛЕЙ МНОЖЕСТВЕННОГО ВЫБОРА*

А.А. САНИНА

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант. E-mail: anastas.sanina@gmail.com

В данной статье рассмотрена задача классификации и некоторые инструменты для ее решения на примере моделей дискретного выбора. Среди предложенных моделей предпочтение отдаётся логит- и пробит-моделям в связи с их «неприхотливостью» к входным факторам. При этом возникает закономерный вопрос о возможности введения новой модели, в основе которой будет лежать некоторая функция, отличная от логистической – для логит-модели и нормальной – для пробит-модели соответственно. В разделе «Постановка задачи и методы решения» подробно описывается математическая формулировка и приводятся пояснения, касающиеся возможности введения новой модели, а также обозначены существующие для этого ограничения. Кроме того, описывается разработанный новый метод оценивания параметров классифицирующей функции, основанный на применении нового распределения. В качестве нового распределения вводится закон Лапласа с неизвестными параметрами. Новая процедура классификации заключается в решении двойной задачи оптимизации: минимизации функции правдоподобия при подборе оптимальных коэффициентов для классифицирующей функции и минимизации значения величины ошибки классификации путем варьирования параметров выбранного распределения. Чтобы сделать исследования более полными, вычислительные эксперименты проводились при различных объемах выборок и переменных для факторов, распределенных согласно стандартному нормальному закону, несимметричному закону на примере экспоненциального распределения, а также распределениям с тяжелыми и легкими хвостами на примере двустороннего экспоненциального закона при различных значениях параметра формы. Полученные результаты свидетельствуют об эффективности предложенной процедуры. Особенно хорошо это иллюстрируют тесты на расширенной модели (с большим количеством переменных). В заключении указаны возможные перспективы развития работы: в связи с тем, что предложенный метод оказался «жизнеспособным», в дальнейшем можно исследовать величину ошибки классификации, выбирая для построения модели любые другие распределения при соблюдении некоторых условий. Немаловажно, что усовершенствованный метод решения за-

* Статья получена 23 января 2015 г.

Работа выполнена при поддержке Министерства образования и науки Российской Федерации, проект 2.541.2014К.

дач классификации дает значительное улучшение качества классификации существующих процедур, а соответственно, может успешно применяться на практике.

Ключевые слова: дискриминантный анализ, логит-модель, пробит-модель, функция правдоподобия, задача классификации, факторы, бинарная зависимая переменная, процедура оптимизации

DOI: 10.17212/2307-6879-2015-1-23-32

ВВЕДЕНИЕ

Задача классификации (или задача принятия решения) встречается практически во всех сферах человеческой деятельности. В настоящее время для ее решения применяются математические модели дискретного выбора: логит- и пробит-модели, частным случаем которых выступает модель дискриминантного анализа [3, 5–13]. Вполне логично, что в какой-то момент перед исследователем встанет закономерный вопрос: какую из моделей предпочесть для решения задачи? Главным критерием отбора модели будем считать «неприхотливость» модели к входным данным.

Модель дискриминантного анализа в этом случае, очевидно, проигрывает [6–10]. Требование выполнения основных предположений дискриминантного анализа, таких как непрерывность, независимость и нормальное распределение факторов, делает модель «нежизненной» в реальных условиях [1, 2].

Логит- и пробит-модели менее требовательны к входным данным, а следовательно, более гибкие [1, 2]. Кроме того, зная, что в основе рассматриваемых моделей лежат логистическое и нормальное распределения соответственно, разумно задуматься над возможностью построения модели с каким-либо другим распределением в основе. В данной работе проводится исследование новой модели с точки зрения качества классификации и сравнение полученных результатов с работой уже известных моделей.

1. ПОСТАНОВКА ЗАДАЧИ И МЕТОДЫ РЕШЕНИЯ

Пусть $z_i = \theta x_i^T = \theta_1 x_{i1} + ... + \theta_n x_{in}$ – модель линейной регрессии, описывающая i-е наблюдение из m, где x_i – вектор значений входных факторов для i-го наблюдения, $x_{ij} \in R$ – значение j-го фактора для i-го наблюдения; $i = \overline{1,m}$, $j = \overline{1,n}$, $\theta = (\theta_1, \theta_2, ..., \theta_n)$ – коэффициенты регрессии.

Зависимая переменная y принимает одно из двух значений: 0 или 1 в зависимости наступления / ненаступления некоторого события.

Основное уравнение модели записывается в виде

$$P\{y=1|\, x_i\} = F(z),$$

где $F(z)$ – некоторая функция распределения, описывающая вероятность возникновения указанного события от входных факторов.

Оценивание параметров $\theta_1, \theta_2, ..., \theta_n$ проводится по набору значений независимых переменных и соответствующих им значений зависимой переменной y. Обычно для этого используется метод максимального правдоподобия, согласно которому оцениваются параметры θ, максимизирующие значение функции правдоподобия. Однако на практике принято использовать эквивалентное логарифмированное выражение для функции правдоподобия:

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

Несложно заметить, что теоретически в качестве функции $F(z)$ может быть взята любая функция распределения, принимающая ненулевые и неединичные значения на всей области определения аргумента. Традиционно в качестве $F(z)$ выбирается логистическое или нормальное распределение для логит- и пробит-моделей соответственно. В данной работе предлагается альтернативный вариант – распределение Лапласа [4]:

$$F(z) = \begin{cases} \dfrac{1}{2}\exp^{\alpha(z-\beta)}, & x \le \beta, \\ 1-\dfrac{1}{2}\exp^{-\alpha(z-\beta)}, & x > \beta, \end{cases}$$

где $\alpha > 0$, $-\infty < \beta < \infty$ будем считать неизвестными.

Пусть Err – величина ошибки классификации (доля неверно классифицированных наблюдений) в результате применения какой-либо модели. Так как функция распределения зависит от параметров, их можно подобрать особым образом, минимизируя данную ошибку:

$$Err = \arg\min_{(\alpha,\beta)}\left(\arg\max_{\theta}(\ln L(\theta))\right). \qquad (1)$$

Следует отметить, что возможны случаи, когда рассмотренные в статье модели не будут работать вовсе в связи с тем, что при определенных значениях факторов и коэффициентов аргумент функции $F(z)$ может оказаться «слишком большим» или наоборот. Функция примет свои экстремальные значения, которые просто «сломают» функцию правдоподобия. В этом случае рекомендуется провести предварительную нормировку входных факторов.

Далее рассмотрим точность классификации в различных условиях.

2. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Проведенные вычислительные эксперименты выполнялись при следующих условиях: факторы – независимые переменные и распределены по непрерывным законам (нормальный – *N*, экспоненциальный – *Exp*, двусторонний экспоненциальный с тяжелыми и легкими хвостами: *DE*(*0.5*), *DE*(*8*)). Выходная переменная – бинарная величина. Количество наблюдений, соответствующее значению $y = 0$, равно n_1, $y = 1$ – соответственно n_2 ($n_1 = n_2$). Общее количество наблюдений *m* = 50, 100, 200, 500.

В табл. 1 и 2 приведены значения показателя *Err* при решении задачи классификации по оцененным значениям параметров θ и параметров распределения Лапласа. Обозначения, принятые в таблицах: Logit – при построении модели использована логит-функция; Probit – при построении модели использована функция нормального распределения; Laplas1 – при построении модели использована функция распределения Лапласа при фиксированных значениях параметров $\alpha = 1$, $\beta = 0$; Laplas2 – решение задачи (1).

Таблица 1

Значения показателя *Err* для модели с одной переменной

Закон	*m*	Logit	Probit	Laplas 1	Laplas 2	Laplas 1– Laplas 2	Laplas 2– Logit
N	50	0.00232	0.00892	0.00056	0.00048	1.16667	0.00008
	100	0.00388	0.00858	0.00158	0.00158	1.00000	0.00000
	200	0.00353	0.00889	0.00179	0.00169	1.05917	0.00010
	500	0.00308	0.00710	0.00188	0.00176	1.06576	0.00012

Окончание табл. 1

Закон	*m*	Logit	Probit	Laplas 1	Laplas 2	Laplas 1–Laplas 2	Laplas 2–Logit
Exp	50	0.01160	0.05412	0.00428	0.00420	1.01905	0.00008
	100	0.01570	0.06584	0.00656	0.00638	1.02821	0.00018
	200	0.01811	0.07546	0.00716	0.00711	1.00703	0.00005
	500	0.02647	0.08789	0.01058	0.01009	1.04917	0.00050
DE(0.5)	50	0.01676	0.12796	0.00856	0.00712	1.20225	0.00144
	100	0.02894	0.16866	0.01428	0.01306	1.09342	0.00122
	200	0.04441	0.19095	0.02049	0.01960	1.04541	0.00089
	500	0.07328	0.22577	0.04029	0.03739	1.07767	0.00290
DE(8)	50	0.00228	0.00732	0.00048	0.00024	2.00000	0.00024
	100	0.00374	0.00810	0.00128	0.00096	1.33333	0.00032
	200	0.00305	0.00699	0.00152	0.00144	1.05556	0.00008
	500	0.00270	0.00618	0.00165	0.00144	1.14444	0.00021

Таблица 2

Значения показателя Err для модели с пятью переменными

Закон	*m*	Logit	Probit	Laplas 1	Laplas 2	Laplas 1–Laplas 2	Laplas 2–Logit
N	50	0.00848	0.02480	0.00372	0.00128	2.90625	0.00244
	100	0.01362	0.03116	0.00728	0.00394	1.84772	0.00334

Окончание табл. 2

Закон	*m*	Logit	Probit	Laplas 1	Laplas 2	Laplas 1– Laplas 2	Laplas 2– Logit
	200	0.01789	0.03817	0.01114	0.00414	2.69082	0.00700
	500	0.01931	0.04104	0.01429	0.00338	4.22840	0.01091
Exp	50	0.01904	0.04724	0.00752	0.00332	2.26506	0.00420
	100	0.02378	0.05626	0.01452	0.00674	2.15430	0.00778
	200	0.02751	0.06323	0.01700	0.00612	2.77778	0.01088
	500	0.03171	0.07398	0.02371	0.00579	4.09323	0.01792
DE(0.5)	50	0.20836	0.22132	0.18424	0.07524	2.44870	0.10900
	100	0.24240	0.26514	0.23080	0.09408	2.45323	0.13672
	200	0.26507	0.28687	0.24737	0.10793	2.29195	0.13944
	500	0.28299	0.31162	0.26536	0.12878	2.06060	0.13658
DE(8)	50	0.09716	0.09456	0.07984	0.02008	3.97610	0.05976
	100	0.11168	0.10968	0.09786	0.01608	6.08582	0.08178
	200	0.11165	0.11100	0.10889	0.01711	6.36411	0.09178
	500	0.11548	0.11564	0.11378	0.01696	6.70738	0.09682

Из обеих таблиц видно, что с точки зрения качества классификации на всех наборах тестов худший результат показывает пробит-модель. При сравнении логит-модели и модели со стандартным распределением Лапласа несложно заметить, что предложенная модель оказывается стабильно лучше. Дополнительная процедура подбора параметров семейства распределения улучшает и без того неплохие значения показателя *Err* до семи раз. Особенно хорошо это видно в случае распределения данных по закону с тяжелыми и легкими хвостами. В процентном соотношении это составляет до 14 %.

При «расширении модели» (увеличении количества независимых переменных с одной до пяти) преимущество новой процедуры (1) становится еще более очевидным, о чем свидетельствуют значения показателя *Err* в табл. 2.

ЗАКЛЮЧЕНИЕ

Проведенные исследования показали, что предложенный в статье метод является «жизнеспособным», поэтому для построения модели допускается использование любого распределения при соблюдении некоторых условий, что дает возможность расширять исследования в этом направлении.

Следует также отметить, что усовершенствованный метод решения задач классификации, заключающийся в варьировании параметров распределения модели, дает значительное улучшение качества классификации существующих процедур и может успешно применяться на практике.

СПИСОК ЛИТЕРАТУРЫ

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

2. *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.

3. *Kropko J.* Choosing between multinomial logit and multinomial probit odels 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.

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

5. *Малхотра Н.К.* Маркетинговые исследования: практическое руководство: пер. с англ. – 3-е изд. – М.: Вильямс, 2002. – 957 с. + Прил. (1 CD-ROM).

6. Электронный учебник по статистике // StatSoft: [мультимедийный портал компьютерной аналитики]. – Москва, 2012. – URL: http://www.statsoft.ru/home/textbook/default.htm (дата обращения: 02.02.2015).

7. *Цильковский И.А., Волкова В.М.* Методы анализа знаний и данных: конспект лекций. – Новосибирск: Изд-во НГТУ, 2010. – 68 с.

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

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

10. *Rencher A.C.* Methods of multivariate analysis. – New York: John Wiley & Sons, 2002. – 727 p.

11. 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.

12. *Кендалл М.Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды: пер. с англ. – М.: Наука, главная редакция физико-математической литературы, 1976. – 736 с.

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

Санина Анастасия Алексеевна, аспирант кафедры ТПИ НГТУ по направлению 09.06.01 «Информатика и вычислительная техника». Основное направление научных исследований – разработка и усовершенствование алгоритмов классификации. E-mail: anastas.sanina@gmail.com.

Improving the quality classification using multiple choice linear models*

A.A. Sanina

Novosibirsk State Technical University, 20 prospect Karla Marksa, Novosibirsk, 630073, Russian Federation, post-graduate student. E-mail: anastas.sanina@gmail.com

In this article we consider a classification problem and some methods for its solving based on the discrete choice models. Logit and Probit Models have been preferred to Discriminant Function Model because they less depend on the input factors. So, the question arises is it possible to introduce a new model based on a function which differs from the logit function for the Logit Model and the normal function for Probit Model respectively. In the "Problem Statement and Solution Methods" section it is described the mathematical model and also illustrated the possibility of introduction of a new model as it is presented existing restrictions preventing this action. Moreover, it is also written about a new method for parameters estimation for the classification function. The method is based on applying a new statistical distribution. The Laplace Law is introduced as a new distribution with unknown parameters. The new classification procedure is to solve the dual optimization problem. They are minimization of the likelihood function with the optimal coefficients fitting for a classification function and

* *Received 23 Janvuary 2015.*

The work was supported by the Ministry of education and science of the Russian Federation, project 2.541.2014K.

minimization of the classification error magnitude by varying the parameters value of the selected distribution. In order to make the study more comprehensive the computational experiments were performed with different sample sizes and varied number of income variables and the factors were distributed according to the standard normal law, asymmetric law based on the exponential distribution, as well as the distributions with heavy and light tails based on the double exponential law with the varied shape parameter value. The obtained results show the effectiveness of the proposed procedure. This is particularly well seen from the tests with the extended model (i.e., the model with many income variables). In "Conclusion" section, possible ways of further development the work have been 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 existing procedures, so it can be successfully applied in practice.

Keywords: Discriminant Function Analysis, Logit model, Probit Model, likelihood function, classification problem, factors, two-valued dependent variable, optimization procedure

DOI: 10.17212/2307-6879-2015-1-23-32

REFERENCES

1. James Press S., Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the America Statistical Assotiation*, 1978, vol. 73, iss. 364, pp. 699–705. doi: 10.1080/01621459.1978.10480080

2. 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.

3. Kropko J. *Choosing between multinomial logit and multinomial probit odels 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.

4. 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.

5. 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).

6. *Elektronnyi uchebnik po statistike* [Electronic textbook statistically]. StatSoft. Moscow, 2005. Available at: http://www.statsoft.ru/home/textbook/default.htm (accessed 02.02.2015)

7. Tsil'kovskii I.A., Volkova V.M. *Metody analiza znanii i dannykh* [Methods of the analysis of knowledge and data]. Novosibirsk, NSTU Publ., 2010. 68 p.

8. 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.).

9. Karimov R.N. *Osnovy diskriminantnogo analiza* [Fundamentals of discriminant analysis]. Saratov, SSTU Publ., 2002. 108 p.

10. Rencher A.C. *Methods of multivariate analysis*. New York, John Wiley & Sons, 2002. 727 p.

11. Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: klassifikatsiya i snizhenie razmernosti*. Moscow, Finansy i statistika Publ., 1989. 607 p.

12. 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.).

13. 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.