УДК 0.04:519.584:57.041

СИНТЕЗ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ НА ПРИНЦИПАХ САМООРГАНИЗАЦИИ МОДЕЛЕЙ

Е.А. Настенко, А.Л. Бойко, Е.К. Носовец, К.И.Тепляков, В.А. Павлов

Национальный технический университет Украины «КПИ»

В работе решается задача структурно-параметрического синтеза модели логистической регрессии. Предложенный алгоритм осуществляет автоматическую оптимизацию параметров шагового алгоритма многомерной логистической регрессии на принципах самоорганизации моделей. Оптимизация параметров осуществляется с помощью предложенного внешнего критерия баланса, отражающего точность классификации на обучающей и проверочных выборках, с одной стороны, и требование к балансу качества распознавания в каждом классе, с другой. Рассмотрен пример моделирования классификатора функциональных состояний сердечнососудистой системы человека. Сравнение результатов моделирования стандартным и предложенным алгоритмами показало преимущество последнего на экзаменационной выборке данных.

Ключевые слова: логистическая регрессия, шаговый алгоритм, оптимизация параметров, принципы самоорганизации моделей, внешний критерий, функциональное состояние, сердечнососудистая система.

У роботі вирішується завдання структурно-параметричного синтезу моделі логістичної регресії. Запропонований алгоритм здійснює автоматичну оптимізацію параметрів крокового алгоритму багатовимірної логістичної регресії на принципах самоорганізації моделей. Оптимізація параметрів здійснюється за допомогою запропонованого зовнішнього критерію балансу, який відображає точність класифікації на навчальній та перевірочних вибірках, з одного боку, та вимогу до балансу якості розпізнавання в кожному класі, з іншого. Розглянуто приклад моделювання класифікатора функціональних станів серцево-судинної системи людини. Порівняння результатів моделювання за стандартним та запропонованим алгоритмами показало перевагу останнього на екзаменаційній вибірці даних.

Ключові слова: логістична регресія, кроковий алгоритм, оптимізація параметрів, принципи самоорганізації, зовнішній критерій, функціональний стан, серцево-судинна система.

Введение

Классические алгоритмы шаговой многомерной регрессии в задачах аппроксимации и классификации [1] реализуют процедуру включения и исключения аргументов в/из модели, исходя из заданных параметров: пороговых значений вероятности ошибок 1-го и 2-го родов или соответствующих пороговых значений критерия Фишера — $F_{\text{вкл}}$ и $F_{\text{искл}}$. При этом задание параметров остается за пользователем, что приводит к субъективному характеру получаемой модели: в зависимости от уровня, типа распределения и точек приложения шумов в данных, при одних и тех же значениях уровней значимости можем получить как переобученную модель, так и модель недостаточной сложности. Принципы решения поставленной проблемы были разработаны в теории самоорганизации [2, 3] на основе введения в алгоритм структурно-параметрического синтеза внешних

критериев — критериев регулярности и несмещенности [2], в методах индуктивного моделирования — критерии Акаике [4], Шварца [5], Маллоуза [6], на основе которых с успехом решается проблема единственности структуры в задачах аппроксимации и прогноза. В настоящей работе рассматривается возможность автоматической оптимизации на принципах самоорганизации параметров шагового алгоритма многомерной регрессии на примере синтеза логистической модели.

Цель работы

Повышение качества классификации шаговых логистических регрессионных моделей за счет автоматической оптимизации параметров алгоритма многомерной бинарной логистической регрессии.

Постановка задачи

Задана матрица входных наблюдений $x \in R^M$ и вектор зависимой переменной $Y \in \{0,1\}$

$$\begin{pmatrix} x_{11} & \dots & x_{1M} & y_1 \\ x_{21} & \dots & x_{2M} & y_2 \\ \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nM} & y_n \end{pmatrix},$$

где n — число наблюдений, M — число переменных, из которых необходимо выбрать наилучших m объясняющих аргументов в модель.

Необходимо построить алгоритм структурно — параметрического синтеза модели оптимальной структуры:

$$Y = \frac{1}{1+\exp^{-z}}, \quad z = b_0 + \sum_{j=1}^{m} b_j x_{i_j} \text{ , где } \quad m < M. \tag{1}$$

Оптимальность структуры модели понимается в смысле достижения наилучшего значения рационально выбранного внешнего критерия на тестовой выборке данных.

Модификация классического алгоритма шаговой регрессии на принципах самоорганизации моделей

В работе предлагается модифицировать на принципах самоорганизации моделей классический алгоритм шаговой регрессии типа Stepwise [6]. Данный алгоритм выбран по следующим причинам.

1. Stepwise реализует наиболее распространенный метод структурно-параметрического синтеза, поэтому его эффективная модификация будет легко воспринята пользователями программного продукта.

2. Метод является одним из первых средств индуктивного моделирования, где используется явный штраф за сложность модели (выражение $d = n - m - 2$ в статистике Фишера при оценке эффективности вводимого/выводимого аргумента). Недостаток такого механизма — крайне низкая чувствительность штрафа при $n >> m$ (n — количество точек, m — сложность модели), что, как правило, и имеет место в практических задачах. В работе предлагается увеличить точность модели на "свежих" (проверочной, экзаменационной) выборках за счет использования принципов самоорганизации моделей [2]: выделение проверочной выборки, используемой для расчета внешнего критерия, осуществление селекции моделей (аргументов) в соответствии с механизмом неявного штрафа, введение внешнего критерия оценки структуры модели.

3. Алгоритм Stepwise не является "жадным", как подавляющее число алгоритмов метода группового учета аргументов [3], реализующих принципы самоорганизации моделей. Это позволяет предполагать высокую эффективность его модификации.

Как отмечено выше, основной проблемой классической процедуры Stepwise есть "ручное" задание оператором моделирования пороговых значений критерия F (или уровней значимости α), используемых при включении или исключении аргумента модели. В работе предлагается задаться сеткой параметров $\alpha_{вкл}$ и $\alpha_{искл}$ и, включив в шаговую процедуру Stepwise расчет внешнего критерия, выбирать модели при тех $\alpha_{вкл}$ и $\alpha_{искл}$, при которых будут получены наилучшие значения внешнего критерия. Внешний критерий должен включать результаты классификации на проверочной выборке данных. Окончательный вариант модели может быть получен ее пересчетом уже на полной выборке наблюдений, при найденных выше оптимальных значениях вероятностей ошибок первого и второго родов $\alpha_{вкл}$ и $\alpha_{искл}$.

Этапы предлагаемого в работе алгоритма выглядят следующим образом.

1. Вводится расширенная (нелинейными аргументами) матрица переменных x.

2. Выравнивается количество наблюдений в классах. Этап введен ввиду несбалансированности результатов классификации логистической моделью, полученной стандартной процедурой итерационного взвешенного метода наименьших квадратов (МНК), при неравенстве наблюдений в классах.

3. Полная выборка наблюдений W делится на обучающую (А), проверочную (В) и экзамен (С) в заданном соотношении случайным образом.

4. Задается текущее значение $\alpha_{вкл}$ и $\alpha_{искл}$ сетки оптимизируемых параметров алгоритма.

5. Для текущего значения параметров алгоритма проводится модифицированная шаговая процедура с определением структуры, коэффициентов логистической модели и расчетом значения внешнего критерия $I_{вн}$. Решением шаговой процедуры считается модель для которой получен максимум внешнего критерия. Структура, коэффициенты и значение внешнего критерия $I_{вн}$ запоминаются.

6. Если все значения сетки перебраны, переходим к п.7, если нет —

задается следующее значение сетки оптимизируемых параметров алгоритма, переход к п. 4.

7. Из полученных моделей выбирается та, для которой получено наилучшее значение внешнего критерия.

8. Окончательной оценкой качества полученной модели вида (1) считаем качество классификации на экзаменационной выборке.

9. Осуществляем пересчет модели на полной выборке W для найденных оптимальных параметров $\alpha_{искл}$ и $\alpha_{вкл}$.

Ниже более подробно описаны отдельные этапы алгоритма:

Этап 3: выборка наблюдений W делится на обучающую A, проверочную B и экзамен C в заданном соотношении случайным образом. Расчет коэффициентов логистических моделей внутри шаговой процедуры будет проводиться на обучающей выборке, значение внешнего критерия для выбора структуры модели будет оцениваться как сбалансированная на классах точность классификации на обучающей и проверочной выборках, результирующая оценка — точность классификации на экзаменационной выборке.

Этап 5: модифицированная шаговая процедура для фиксированных значений $\alpha_{вкл}$ и $\alpha_{искл}$ состоит из нескольких шагов.

1. Включение предиктора в модель:

1.1. Проводим F-тест, сравнивая модель, полученную на предыдущей итерации, с моделью, включающую предиктор (переменную) x_i для каждого предиктора, еще не включенного в модель:

$$F_i = \frac{SSR_{\text{prev}+x_i} - SSR_{\text{prev}}}{MSR_{\text{prev}+x_i}}, \tag{2}$$

$$\text{где } MSR_{\text{prev}+x_i} = \frac{SSR_{\text{prev}+x_i}}{d}, \quad d = n - m - 2, \tag{3}$$

$$SSR = \sum_{i=1}^{n} \overline{Y}_i - Y, \tag{4}$$

$i = 1,...,k_1$, k_1 — количество предикторов претендентов ранее не включенных в модель, Y — табличное значение выходной переменной, Y_i — значение регрессионной модели, n — количество наблюдений в выборке, m — количество переменных в модели, индекс prev — означает модель, полученную на предыдущей итерации.

1.2. Выбираем предиктор с наибольшим значением F. Если уровень значимости α, соответствующий полученному значению критерия F, меньше фиксированного $\alpha_{вкл}$ ($\alpha < \alpha_{вкл}$), то принимается гипотеза о включении предиктора в модель.

2. Исключение предикторов:

2.1. Проводим F-тест, сравнивая текущую модель с моделью, не включающую предиктор x_i для каждого предиктора из модели:

$$F_i = \frac{SSR_{\text{curr}} - SSR_{\text{curr}-x_i}}{MSR_{\text{curr}}} \tag{5}$$

$$MSR_{\text{curr}} = \frac{SSR_{\text{curr}}}{d} \tag{6}$$

где $i = 1,...,k_2$, k_2 — количество предикторов ранее включенных в модель, индекс curr — означает модель, полученную на текущей итерации.

2.2. Если уровень значимости α, соответствующий полученному значению критерия F, превышает фиксированный $\alpha_{\text{искл}}$ ($\alpha > \alpha_{\text{искл}}$), то принимается гипотеза об исключении предиктора из модели. Среди переменных прошедших F-тест для исключения выбираем предиктор с наименьшим значением F.

Для каждой из сравниваемых выше моделей, их коэффициенты считаются по итерационному взвешенному методу наименьших квадратов [8] на обучающей выборке данных.

Рассчитываем и запоминаем значение внешнего критерия $I_{\text{вн}}$ для модели полученной на данной итерации. Если предикторы для включения не исчерпаны, переходим к п. 2.1., если исчерпаны — к п. 2.3.

2.3. Для фиксированного значения параметров сетки $\alpha_{\text{вкл}}$ и $\alpha_{\text{искл}}$ запоминается модель с наилучшим значением внешнего критерия $I_{\text{вн}}$.

Определим далее целесообразную форму внешнего критерия. Отразим в нем требования к точности классификации на обучающей и проверочных выборках, с одной стороны, и требование к балансу качества распознавания в каждом классе, с другой. Тогда внешний критерий (назовем его критерий баланса качества классификации — БКК) имеет вид:

$$I_{вн} = \alpha\Delta^A + (1-\alpha)\Delta^B, \tag{7}$$

где α — коэффициент баланса качества классификации на обучающей и проверочной выборках, Δ^A — критерий качества классификации на обучающей выборке A, Δ^B — критерий качества классификации на проверочной выборке B.

Критерий Δ^A рассчитывается на обучающих точках A для фиксированной структуры и соответствующей подматрицы X:

$$\Delta^A = (1-\gamma) * (\Delta^{A_1} + \Delta^{A_2}) + \gamma * \frac{1}{1 + \left|\Delta^{A_1} - \Delta^{A_2}\right|}. \tag{8}$$

Критерий Δ^B рассчитывается для проверочных точек B и моделей с коэффициентами, найденными на обучающей выборках A:

$$\Delta^B = (1-\gamma) * (\Delta^{B_1} + \Delta^{B_2}) + \gamma * \frac{1}{1 + \left|\Delta^{B_1} - \Delta^{B_2}\right|}. \tag{9}$$

где r — коэффициент баланса между суммарным качеством распознавания по классам и симметричностью (относительно каждого класса) полученного результата;

$$\Delta^{A_1} = \frac{n^{A_1^*}}{n^{A/2}} = \frac{2n^{A_1^*}}{n^A}, \qquad \Delta^{A_2} = \frac{n^{A_2^*}}{n^{A/2}} = \frac{2n^{A_2^*}}{n^A}, \tag{10}$$

$$\Delta^{B_1} = \frac{n^{B_1^*}}{n^{B/2}} = \frac{2n^{B_1^*}}{n^B}, \qquad \Delta^{B_2} = \frac{n^{B_2^*}}{n^{B/2}} = \frac{2n^{B_2^*}}{n^B}, \tag{11}$$

где $n^{A_1^*}$, $n^{B_1^*}$ — количество правильно распознанных объектов 1-го класса; $n^{A_2^*}$, $n^{B_2^*}$ — количество правильно распознанных объектов 2 класса; n^A, n^B — количество объектов в выборке.

Сравнительная оценка работы алгоритмов при решении задачи построения классификатора функциональных состояний сердечнососудистой системы

Для сравнительной оценки базового и предложенного алгоритмов решалась следующая задача: получить классифицирующую функцию, выделяющую группу испытуемых с некоторым определенным функциональным состоянием сердечно-сосудистой системы (23 наблюдения, значение классифицирующей переменной "1") от остальной выборки испытуемых (138 наблюдений, значение классифицирующей переменной "0"). Для расчета были взяты данные, полученные в лаборатории функциональной диагностики кафедры Физического воспитания НТУУ «КПИ». Выборка содержит 161 наблюдение и 22 переменных, характеризующих антропометрические параметры и психо-физиологические тесты, проведённые для студентов 2-го курса университета.

Рассмотренный выше алгоритм реализован в программной среде R. Ниже приведены результаты расчета логистической модели и сравнения качества классификации стандартным алгоритмом логистической регрессии glm в программной среде R с предложенной в работе версией алгоритма шаговой регрессии на принципах самоорганизации моделей с оптимизацией параметров $\alpha_{вкл}$ и $\alpha_{искл}$ (АШРО).

Выборка была разбита на обучающую, проверочную и экзаменационную в пропорции — 55/30/15. Модели логистической регрессии сравниваемыми алгоритмами считались на рабочей выборке (обучающая + проверочная), а окончательная оценка алгоритмов осуществлялась на экзаменационной выборке.

На рис. 1 приведен график чувствительности (Sensitivity) и специфичности (Specificity), а также ROC кривая результатов работы стандартного алгоритма glm логистической регрессии в R на экзаменационной выборке.

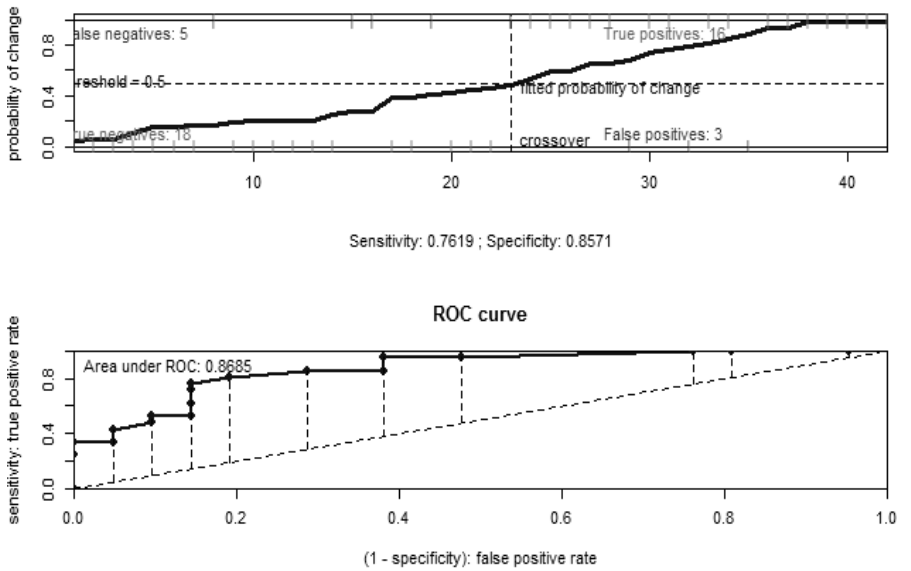

Рис. 1. Результаты работы стандартного алгоритма glm

Стандартный алгоритм показал на экзаменационной выборке качество классификации — 81 %, область под ROC-кривой — 0,87.

На рис. 2 приведены графики чувствительности (Sensitivity) и специфичности (Specificity) и ROC-кривая для АШРО рассчитанные на экзаменационной выборке данных. Качество классификации данного алгоритма — 90,5 %, область под ROC — кривой — 0,97.

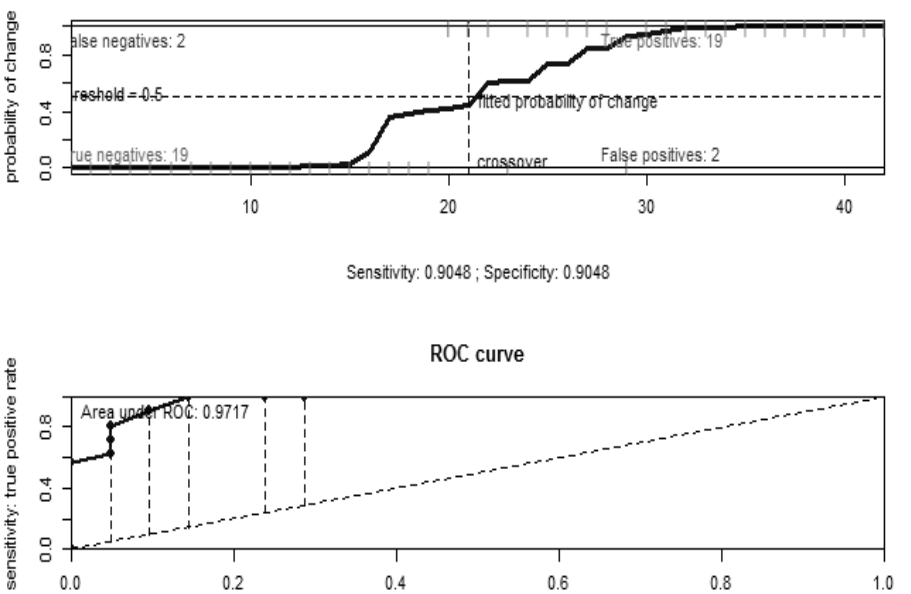

Рис. 2. Результаты работы АШРО алгоритма

Выводы

В решении задачи классификации функциональных состояний сердечнососудистой системы, предложен алгоритм синтеза логистической регрессии на принципах самоорганизации моделей, который при сравнении со стандартным шаговым алгоритмом логистической регрессии, показал улучшение результатов классификации экзаменационной выборки на 10 %,. Эффект получен за счет оптимизации параметров шагового алгоритма в соответствии с предложенным в работе внешним критерием, Критерий отражает требования к балансу точности классификации на обучающей и проверочных выборках, с одной стороны, и балансу качества распознавания в каждом классе с другой.

1. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей — М.: ВЦ РАН. — 2010. — С. 20–23.
2. Ивахненко А.Г., Степашко В. С. Помехоустойчивость моделирования — Киев.: Наук.думка. — 1985. —216 С.
3. Ивахненко А.Г., Мюллер Й.А. Самоорганизация прогнозирующих моделей — К.: Техніка ; Берлин : Феб Ферлаг техник. — 1984. — 223 С.
4. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. — 1974. — Vol. 19. — P. 716–723.
5. Schwarz E. Estimating the dimension of a model // Annals of Statistics. — 1978. — Vol. 6, № 2. — P. 461–464.
6. Mallows C.L. Some Comments on CP / C.L. Mallows // Technometrics — 1973. — Vol. 15, № 4 — P. 661–675.
7. Efroymson M.A. Multiple regression analysis //Mathematical Methods for Digital Computers, — 1960.
8. Green P.G. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives (with discussions) // Journal of the Royal Statistical Society, Series — 1984. — B 46, — P. 149–192.

UDC 0.04:519.584:57.041

SYNTHESIS OF LOGISITIC REGRESSION, BASED ON SELF-ORGANISATION PRINCIPLES OF MODELS

E.A. Nastenko, A.L. Boyko, O.K. Nosovets, K.I. Teplyakov, V.A. Pavlov

National Technical University of Ukraine "Kiev Polytechnical Institute" (Kiev)

Introduction. Requirements for modeling algorithms and their implementations varies depending upon the desired properties of the models, which has to be received in restrictions on the available computational resources. Examples of desired properties — accuracy, efficiency ratings, the lowest sensitivity to a change in the data of the model error, variance estimation of parameters, p values etc. Depending on the specific use of models, those or other criteria are taken as a basis for designing specific algorithm simulation. However, choice of the solution of resulting model is usually left to the user. This article considers the possibility of stepwise regression algorithm's automatic optimization of parameters that is based on principles of self-organization on an example of the synthesis of the logistic model.

The purpose of this article is the improvment the quality of logistic regression

classification models due to automatic optimization multivariate binary logistic regression algorithm parameters.

Results. The essence of the modification of stepwise logistic regression standard algorithm: defines p_{enter} , p_{leave} grid for each combination of the thresholds calculates stepwise logistic algorithm and the corresponding value of the external criteria. Proposed external criteria reflects the classification accuracy on the training and test datasets, on the one hand, and the requirement to balance the quality of recognition in each class on the other. The stated procedure is repeated for the next value of the grid parameters of the algorithm. Final evaluation of the model is given in the exam sample data. For logistic model calculation and quality's comparison of classification between standard logistic regression (glm function in R software) and proposed version of modified stepwise algorithm were taken data obtained in the laboratory of functional diagnostics at Department of Physical Education NTUU "KPI". The purpose of the example is to get a classifying function, of group of subjects with certain states of the cardiovascular system from the rest of the test sample. Standard algorithm demonstrated on examination sample classification quality — 81%, the area under the ROC — curve — 0.8685. Graphs of sensitivity and specificity, and ROC curve for modified algorithm showed the results: quality of the classification algorithm — 90.5 %, area under the ROC — curve — 0.9717.

Conclusions. Article proposes stepwise logistic regression based on the principles of self-organization synthesis algorithm. In order to optimize the parameters of the algorithm proposed by external criterion, which reflects the classification accuracy on the training and test samples and requirement to balance the quality of recognition in each class the effect was received. For the aboved example the classification of functional states of the cardiovascular system in comparison of the standard stepwise algorithm with the proposed algorithm has shown classification quality improvement on 10 % on examination sample.

Keywords: logistic regression, stepwise regression, self-organization's principles.

1. Strighov V., Krimova E., Selection methods of regression models — Moscow: CC RAS — 2010. — 45 p. (in Russian).
2. Ivakhnenko A., Stepashko V. Noisestability modelling — Kiev: «Nauk.dumka». — 1985, — 216 p. (in Russian).
3. Ivakhnenko A. Muller J. Self-organization of predictive models — Kiev: Technic. — 1984, — 223 p. (in Russian).
4. Akaike H.A new look at the statistical model identification // IEEE Transactions on Automatic Control — 1974. — Vol. 19. — P.716–723.
5. Schwarz E. Estimating the dimension of a model // Annals of Statistics — 1978. — Vol. 6. — № 2. — P. 461–464.
6. Mallows C.L. Some Comments on CP//Technometrics — 1973. — Vol. 15. — № 4. — P. 661–675.
7. Efroymson M.A. Multiple regression analysis // Mathematical Methods for Digital Computers — 1960.
8. Green P.G. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives (with discussions) // Journal of the Royal Statistical Society, Series — 1984. — B 46. — P. 149–192.

Получено 15.06.2015