Известия Тульского государственного университета
Естественные науки. 2012. Вып. 1. С. 153–162

Информатика

УДК 519.252

Многоклассовая логистическая регрессия для прогноза вероятности наступления инфаркта *

А. П. Мотренко, В. В. Стрижов

Аннотация. В работе описан алгоритм классификации четырех групп пациентов: перенесших инфаркт, имеющих предрасположенность к инфаркту и здоровых пациентов двух типов. Признаками для определения состояния пациента служат измерения концентрации белков в крови. Работа посвящена прогнозу вероятности принадлежности пациента к одному из нескольких неупорядоченных классов. Решается задача оценки параметров функции регрессии и выбора признаков при многоклассовой классификации. Классификация выполняется по всевозможным парам групп.

Ключевые слова: логистическая регрессия, многоклассовая классификация, выбор признаков, прогноз предрасположенности к инфаркту.

Введение

Заболевания сердечно-сосудистой системы могут протекать, не проявляясь клинически. Тем не менее, обнаружение нарушений, связанных с работой сердца, по косвенным признакам вполне возможно [1, 2]. В данной работе в качестве признаков (биомаркеров) используются концентрации белков и их соединений, абсорбированные на поверхности кровяных телец. Разделение пациентов на группы по состоянию здоровья приводит к задаче многоклассового прогнозирования. Эта задача сведена к задаче двуклассовой классификации; используется подход «каждая группа против каждой». В этом случае рассматриваются все возможные пары групп пациентов и решается задача вида «к какой из двух данных групп пациент принадлежит с большей вероятностью?». Данный подход принят в связи с относительно небольшим объемом выборки, на которой проводился вычислительный эксперимент.

* Работа выполнена при финансовой поддержке РФФИ (проект № 10-07-00422).

Для каждой пары групп решается задача логистической регрессии [3], в основе которой лежит предположение о биномиальном распределении независимой переменной, и оцениваются параметры функции регрессии [4, 5].

Предполагается, что число измеряемых признаков избыточно; требуется отыскать оптимальный набор признаков, эффективно разделяющий классы. Признаки в логистической регрессии как правило выбираются с помощью шаговой регрессии [6, 7]. В данной работе используется полный перебор, т.к. он дает экспертам гарантию, что рассмотрены все возможные сочетания признаков при выборе модели. При этом экспертами вводились ограничения на сложность модели. Задача выбора признаков поставлена с использованием площади под ROC-кривой [8] в качестве внешней функции ошибки.

Задача классификации сопряжена с оценкой минимального объема выборки, достаточного для проведения классификации. Для этого используются метод доверительных интервалов [9], метод скользящего контроля [10], сравнение предполагаемых распределений на различных подвыборках [11].

При проведении вычислительного эксперимента и прогноза вероятности наступления инфаркта были использованы данные [12], предложенные специалистами парижской лаборатории анализа крови «Иммуноклин».

1. Задача классификации и оценка параметров

Дана выборка $D = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m$, состоящая из m объектов (пациентов), каждый из которых описывается n признаками (биомаркерами) $\mathbf{x}_i \in \mathbb{R}^n$ и принадлежит одному из двух классов $y_i \in \{0, 1\}$. Рассмотрим задачу логистической регрессии. Предполагается, что вектор ответов $\mathbf{y} = [y_1, \ldots, y_m]^T$ — бернуллиевский случайный вектор с независимыми компонентами $y_i \sim \mathcal{B}(p_i, 1 - p_i)$ и плотностью

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{1}$$

Определим функцию ошибки следующим образом:

$$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln p_i + (1 - y_i) \ln (1 - p_i). \tag{2}$$

Другими словами, функция ошибки есть логарифм плотности, или функции правдоподобия, со знаком минус. Требуется оценить вектор параметров $\widehat{\mathbf{w}}$, доставляющий минимум функции ошибки:

$$\widehat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w}). \tag{3}$$

Вероятность принадлежности объекта к одному из двух классов определим как

$$p_i = \frac{1}{1+\exp(-\mathbf{x}_i^T\mathbf{w})} = \sigma(\mathbf{x}_i^T\mathbf{w}) \equiv \sigma_i. \tag{4}$$

Для оценки параметров, воспользовавшись тождеством

$$\frac{d\sigma(\theta)}{d\theta} = \sigma(1-\sigma),$$

вычислим градиент функции $E(\mathbf{w})$:

$$\nabla E(\mathbf{w}) = -\sum_{i=1}^{m}\big(y_i(1-\sigma_i) - (1-y_i)\sigma_i\big)\mathbf{x}_i = \sum_{i=1}^{m}(\sigma_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\sigma} - \mathbf{y}),$$

где вектор $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_m]^T$ и матрица $\mathbf{X} = \big[\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T\big]^T$ состоит из векторов-описаний объектов.

Оценка параметров осуществляется по схеме Ньютона-Рафсона. Введем обозначение $\boldsymbol{\Sigma}$ — диагональная матрица с элементами $\Sigma_{ii} = \sigma_i(1-\sigma_i)$, $i = \\ = 1, \ldots, m$. В качестве начального приближения $\mathbf{w} = [w_1, \ldots, w_n]^T$ вектора $\widehat{\mathbf{w}}$ возьмем

$$w_j = \sum_{i=1}^{m} y_i(1-y_i), \quad j = 1, \ldots, n.$$

Оценка параметров $\mathbf{w}_{k+1}$ логистической регрессии (4) на $k+1$-м шаге итеративного приближения имеет вид

$$\mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\sigma} - \mathbf{y}) = (\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}(\mathbf{X}\mathbf{w}_k - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\sigma} - \mathbf{y})). \tag{5}$$

Процедура оценки параметров повторяется, пока норма разности $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|_2$ не станет достаточно мала.

Алгоритм классификации имеет вид:

$$a(\mathbf{x}) = \operatorname{sign}\big(\sigma(\mathbf{x}, \mathbf{w}) - \sigma_0\big), \tag{6}$$

где σ_0 — задаваемое в (8) пороговое значение функции регрессии (4).

Вычисления качества прогноза. Для контроля за качеством прогноза множество индексов объектов $\mathcal{I}$ разбивается случайным образом на два подмножества, $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$, обучающее и тестовое. Параметры $\mathbf{w}$ оцениваются на подвыборке $D_{\mathcal{L}}$, а качество прогноза вычисляется на подвыборке $D_{\mathcal{T}}$. В данной работе для оценки качества прогноза и для выбора признаков используется один из двух функционалов: площадь AUC под кривой ROC и функционал

$$Q = (1 - \mathrm{TPR})^2 + \mathrm{FPR}^2,$$

где

$$\mathrm{TPR} = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) = 1][y_i = 1]$$

есть доля объектов выборки, правильно классифицированных в пользу данного класса, и

$$\mathrm{FPR} = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i)=1][y_i=0]$$

есть доля ошибочно классифицированных в пользу данного класса объектов выборки. Здесь используется обозначение индикаторной функции:

$$[y=1] = \begin{cases} 1, & y=1; \\ 0, & y\neq 1. \end{cases} \tag{7}$$

Таким образом, алгоритм тем лучше разделяет классы, чем меньше значение функционала Q или чем больше значение площади AUC под кривой ROC. Отложив на графике для каждого значения связанной переменной $\xi \in [0,1]$ по оси абсцисс значения $\mathrm{FPR}(\xi)$, а по оси ординат — $\mathrm{TPR}(\xi)$ получим кривую ROC, каждая точка которой соответствует некоторому значению σ_0.

Отыскание параметра σ_0 алгоритма классификации. В алгоритме (6) используется то значение σ_0, которое соответствует наибольшему расстоянию от отрезка $[(0,0),(1,1)]$, означающего отказ от принятия решения о классификации, до кривой ROC:

$$\widehat{\sigma}_0 = \arg\max_{\xi\in[0,1]} \left\| \big(\mathrm{TPR}(\xi), \mathrm{FPR}(\xi)\big) - (\xi,\xi) \right\|_1 . \tag{8}$$

Последнее выражение включает вычисление значения функционала качества, и как следствие, вычисление выражения (6) и итеративную оценку параметров (5).

2. Выбор признаков в задаче классификации

Введем обозначения $\mathcal{A}$ — некоторое подмножество индексов признаков, $\mathcal{A} \subseteq \mathcal{J} = \{1,\ldots,n\}$ и $\widehat{\mathcal{A}}$ — оптимальный набор индексов. Обозначим $\mathbf{X}_{\mathcal{A}}$ множество столбцов-признаков матрицы $\mathbf{X}$, заданное набором $\mathcal{A}$ и $\mathbf{w}_{\mathcal{A}}$ — соответствующие им параметры. Рассмотрим задачу выбора признаков как задачу максимизации:

$$\widehat{\mathcal{A}} = \arg\max_{\mathcal{A}\subseteq\mathcal{I}} \mathrm{AUC}(\mathcal{A}) \quad \text{при условии} \quad |\mathcal{A}| = \mathrm{const}. \tag{9}$$

В задаче использована площадь под кривой $\mathrm{AUC}(\mathcal{A}) \equiv \mathrm{AUC}(\mathbf{X}_{\mathcal{A}}\widehat{\mathbf{w}}_{\mathcal{A}}, \widehat{\sigma}_0, \mathbf{y})$, значение которой вычислено для набора индексов признаков $\mathcal{A}$, а параметры $\widehat{\mathbf{w}}_{\mathcal{A}}$ и σ_0 получены в результате решения задач (3) и (8).

Набор признаков отыскивается путем полного перебора. Такой подход возможен благодаря сравнительно небольшому количеству признаков в данной задаче и диктуется требованиями экспертов. Запишем выражение для функции регрессии $\sigma(\mathbf{x}_i^T\mathbf{w})$ в виде

$$\mathbf{x}_i^T\mathbf{w} = \alpha_1 x_{i1} w_1 + \alpha_2 x_{i2} w_2 + \ldots + \alpha_n x_{in} w_n,$$

где $\alpha_j \in \{0,1\}$ — структурный параметр. Таким образом, алгоритм нахождения признаков сводится к перебору значений элементов вектора $[\alpha_1,\ldots,\alpha_n]$ структурных параметров

α_1	α_2	$\ldots$	α_n
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	1	1

.

В данном случае не оговаривается необходимость разбиения выборки D на обучающую и тестовую подвыборки, так как эксперты зафиксировали максимальное число признаков при решении задачи: $|\mathcal{A}|$ не должна превышать четырех. Наборы признаков, полученные в результате решения задачи (9), будем называть оптимальным для данной пары классов, а сами признаки — наиболее информативными (рис. 1 и 2).

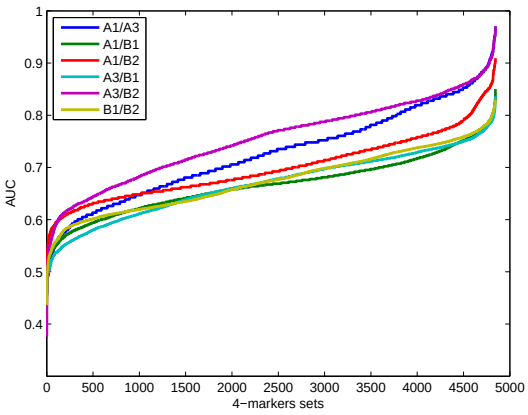

Рис. 1. Вариационный ряд значений функционала для всех наборов признаков фиксированной мощности

3. Прогноз при многоклассовой классификации

Пациенты из исследуемой выборки разделены по состоянию здоровья на четыре группы:

A_1 — пациенты, уже перенесшие инфаркт,

A_3 — пациенты, имеющие предрасположенность к инфаркту,

B_1, B_2 — здоровые пациенты двух типов.

При появлении в выборке нового объекта $\mathbf{x}_{m+1}$, состояние которого необходимо спрогнозировать, выполняется следующая процедура. Для каждой пары групп из шести возможных пар выполняется

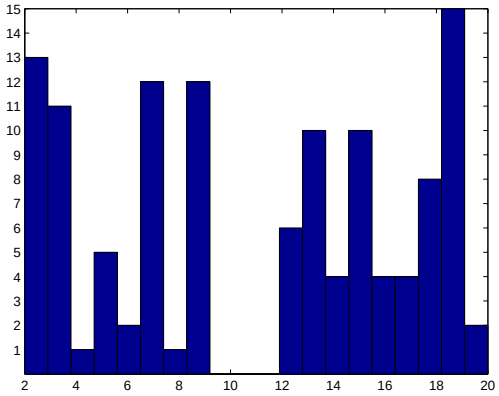

Рис. 2. Число вхождений каждого из двадцати маркеров в набор «K лучших»

классификация (6). Используется оптимальный для соответствующей пары набор признаков (9). В каждом случае алгоритм $a(\mathbf{x}_{m+1})$ возвращает решение о принадлежности объекта к одному из двух рассматриваемых классов и вероятность принадлежности $p_{m+1} = \sigma(\mathbf{x}_{m+1}^T \mathbf{w})$. По этим результатам составляется таблица следующего вида:

	A_1	A_3	B_1	B_2
A_1	–	0	0	1
A_3	1	–	1	1
B_1	1	0	–	0
B_2	0	0	1	–

. (10)

Например, третья строка содержит следующие результаты классификации: объект $\mathbf{x}_{m+1}$ скорее принадлежит классу B_1, чем классу A_1, и скорее принадлежит классам A_3 и B_2, чем классу B_1. Присвоим классам A_1, A_3, B_1, B_2 номера $1, 2, 3, 4$ и введем нижние индексы $a_{lk}(\mathbf{x}) \in \in \{0, 1\}$, $l, k \in \{1, \ldots, 4\}$ для пары классов (l, k). Тогда для рассмотренного примера

$$a_{23}(\mathbf{x}_{m+1}) = 0.$$

При прогнозировании объект относится к тому классу, для которого сумма элементов таблицы по строке наибольшая:

$$\text{class}(\mathbf{x}_{m+1}) = \arg \max_{l \in \{1, \ldots, 4\}} \sum_{k=1}^{4} a_{lk}(\mathbf{x}_{m+1}). \tag{11}$$

Если эта сумма для двух классов совпала, результатом будет классификация (6), полученная для этих двух классов.

4. Результаты вычислительного эксперимента

Вычислительный эксперимент проводился на данных лаборатории анализа крови «Иммуноклин» [12]. Данные содержат измерения концентрации 20-ти белков и их соединений на поверхности кровяных телец 98-ми пациентов четырех классов; в каждом классе примерно равное число пациентов. В табл. 1 приведен список исследуемых биомаркеров с их порядковыми номерами.

Таблица 1

Список исследуемых биомаркеров

1	2	3	4	5	6	7	8	9	10
K	L	K/M	L/M	K/N	K/O	L/O	K/P	L/P	K/Q
11	12	13	14	15	16	17	18	19	20
K/R	L/R	L/R/SA	L/T/SA	L/T/SO	U/V	U/W	U/X	U/Y	U/Z

Построение процедуры прогнозирования выполнялась следующим образом.

(1) Для множества объектов обучающей выборки, принадлежащих каждой паре классов, рассматривались все наборы $\{\mathcal{A}\}$ признаков мощностью 4.

(2) Оценивались параметры $\widehat{\mathbf{w}}, \widehat{\sigma}_0$ алгоритма классификации, соответствующие набору $\mathcal{A}$.

(3) Вычислялось значение функции качества алгоритма классификации AUC на выборке, состоящей из пары классов.

(4) Для каждой пары классов выбраны K наборов признаков, доставляющих большее значение функции качества AUC.

Таблица 2

Результаты выбора признаков

Классы	m_2	m_1	$\mathcal{A}$	AUC($\mathcal{A}$)
$A_1 - A_3$	31	14	[2, 11, 19, 20]	0.953
			[2, 13, 19, 20]	0.953
$A_1 - B_1$	55	14	[3, 13, 18, 19]	0.829
			[12, 13, 15, 19]	0.829
$A_1 - B_1$	55	14	[5, 15, 17, 19]	0.901
			[6, 12, 15, 19]	0.901
$A_3 - B_1$	58	17	[5, 6, 11, 17]	0.814
			[2, 7, 9, 13]	0.829
$A_3 - B_2$	43	17	[2, 3, 5, 9]	0.954
			[2, 3, 9, 19]	0.957
$B_1 - B_2$	67	41	[1, 2, 3, 9]	0.821
			[2, 3, 9, 11]	0.823

В табл. 2 для каждой пары классов указаны наборы маркеров, доставивших наибольшие значения максимизируемому критерию AUC, и сами значения этого критерия. Приведены два набора для каждой пары классов. Для исследования были выбраны пять лучших наборов. Число наборов K задано экспертами, исходя из рис. 1, на котором изображен вариационный ряд (в порядке возрастания) значений AUC, полученных при классификации на различных наборах $\{\mathcal{A}\}$. Предлагается выбирать такое число K, при котором рост графика меняется еще достаточно сильно (справа налево). В данном случае выбрано значение $K = 5$ (табл. 3).

Таблица 3

Число вхождений признаков в K оптимальных наборов для каждой пары классов.

Классы	K	L	K/M	K/N	K/O	L/O	K/P	L/P	K/R
$A_1 - A_3$	0	5	0	0	0	0	0	0	1
$A_1 - B_1$	0	0	1	0	0	0	1	0	0
$A_1 - B_2$	0	0	1	1	1	0	0	1	0
$A_3 - B_1$	0	3	1	3	2	2	0	3	1
$A_3 - B_2$	0	5	4	1	0	0	0	5	0
$B_1 - B_2$	2	5	3	0	0	0	0	3	1

Классы	L/R	L/R/SA	L/T/SA	L/T/SO	U/V	U/W	U/X	U/Y	U/Z
$A_1 - A_3$	0	1	1	0	1	1	0	5	5
$A_1 - B_1$	2	4	0	2	0	0	4	5	1
$A_1 - B_2$	4	0	0	5	0	2	0	5	0
$A_3 - B_1$	0	2	0	0	0	1	1	0	1
$A_3 - B_2$	0	1	0	0	0	1	1	2	0
$B_1 - B_2$	0	1	0	0	0	0	3	2	0

Согласно полученным результатам, итоговый алгоритм прогнозирования имеет следующий вид.

(1) Для нового объекта выполняется классификация на всех K наборах признаков каждой пары классов.

(2) Строится таблица (10), включающая каждую пару классов K раз.

(3) Решается задача прогнозирования (11) с учетом числа наборов K.

Одной из важных практических задач, решаемых в рамках проводимых исследований, является задача снижения стоимости клинического обследования одного пациента, решаемая путем уменьшения числа измеряемых биомаркеров. Предложено измерять только наиболее информативные биомаркеры, выбранные следующим образом. Для j-той пары классов найдено множество оптимальных наборов $\mathcal{S}_j = \bigcup\limits_{i=1}^{K} \mathcal{A}_{j_i}$, $j = 1, \ldots, 6$. Объединив признаки из всех наборов из колонки «$\mathcal{A}$» табл. 2, получим множество наиболее информативных признаков $\bigcup\limits_{j=1}^{6} \mathcal{S}_j$. Для каждого признака подсчитано количество его вхождений в это множество.

Гистограмма на рис. 2 показывает, насколько часто каждый признак входит в K лучших наборов, полученных для каждой пары классов. Таблица 3 показывает число вхождений биомаркера в набор наиболее информативных признаков каждой пары групп пациентов.

Заключение

В работе описан алгоритм прогнозирования вероятности наступления инфаркта пациентов при многоклассовом прогнозировании; описан способ оценки параметров и выбора наиболее информативных признаков. Выборка пациентов разбита на четыре группы относительно наличия нарушений в работе сердечно-сосудистой системы. При этом задача многоклассовой классификации сводилась к двуклассовой путем рассмотрения всевозможных пар групп. Для каждой из таких пар получен оптимальный набор признаков. Получена оценка качества прогнозирования в парах групп на исследуемой выборке.

Список литературы

1. *Azuaje F., Devaux Y., Wagner D.* Computational biology for cardiovascular biomarker discovery // Brief Bioinform. 2009. V.10. №4. P.367–377.
2. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure / Heidecker [et al.] // Circulation. 2008. V.118. №3. P.238–246.
3. *Hosmer D., Lemeshow S.* Applied logistic regression. N. Y.: Wiley, 2000. 375 p.
4. *Bishop C.M.* Pattern recognition and machine learning. Springer, 2006. 738 p.
5. *MacKay D.J.C.* Information theory, inference, and learning algorithms. Cambridge University Press, 2003. 628 p.
6. *Friedman J., Hastie, Tibshirani R.* Additve logistic regression: a statistical way of boosting // The Annals of Statistics. 2000. V.28. №2. P.337–407.
7. *Madigan D., Rideway G.* Discussion of least square regression / В сб. Efron B. [et al.]. Least Angle Regression // The Annals of Statistics. 2004. V.32. №2. P.465–469.
8. *Fawcet T.* ROC graphs: notes and practical considerations for researchers // HP Laboratories. 2004. 38 p.
9. *Реброва О.Ю.* Статистический анализ медицинских данных. Применение прикладного пакета Statistica. М.: Медиасфера, 2002. 312 с.
10. *Bos S.* How to partition examples between cross-validation set and fraining set? / Saitama. Japan: Laboratory for information representation RIKEN. 1995. 4 p.
11. *Perez-Cruz F.* Kullback-Leibler divergence estimation of continuous distributions // IEEE International Symposium on Information Theory, 2008.
12. Standart flow cytometry analysis of nondental patients. Paris: ImmunoClin laboratory. 2007. 1 p.

Стрижов Вадим Викторович (strijov@ccas.ru, http://strijov.com), к.ф.-м.н., н.с., Вычислительный центр Российской Академии Наук, Москва.

Мотренко Анастасия Петровна (pastt.petrovna@gmail.com), студент, Московский физико-технический институт.

Multiclass logistic regression for cardio-vascular disease forecasting

A. P. Motrenko, V. V. Strijov

Abstract. The paper describes an algorithm to classify four groups of patients: a cardio-vascular disease group, a cardio-risk group and two types of healthy groups. The blood-cells protein measurements are the description features for an investigated patient. The paper develops an algorithm to forecast a patient's cardio-vascular disease case as one of four unordered classes. The problem is to estimate the regression parameters and select the most informative features for multi-class classification. During the forecasting all pairs of the classes are considered.

Keywords: logistic regression, multiclass classification, feature selection, cardio-vascular disease forecasting.

Strijov Vadim (strijov@ccas.ru, http://strijov.com), candidate of physical and mathematical sciences, researcher, Computing Center of the Russian Academy of Sciences, Moscow.

Motrenko Anastasiya (pastt.petrovna@gmail.com), student, Moscow Institute of Physics and Technology.

Поступила 01.02.2012