doi: 10.32620/oikit.2019.83.14

УДК 57.087

К.А. Базилевич, Е.С. Меняйлов, С.И. Горанина, К.А. Федулов

Определение вероятности заболевания болезнями сердца на основе методов Data Mining

Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт»

В современном мире специалисты разных областей ищут способы лечения и диагностирования заболеваний. Решение проблемы, которая заключается в ограниченных возможностях своевременной диагностики, лежит в области методов Data Mining.

Первым из методов, рассматриваемых в работе, является метод оценивания параметров логистической регрессии на основе метода оценки шансов и вероятностей. Метод чаще всего используется для выборок с малым количеством параметров. Если в выборке большое количество параметров, этот метод теряет свою точность. Вторым рассматриваемым методом является метод оценки вероятности заболевания с использованием байесовского классификатора. Этот метод более выгоден для применения на выборках с большим количеством параметров, так как метод не теряет своей точности при увеличении количества переменных, однако, несмотря на это, количество признаков должно быть постоянным. В случае с переменным количеством признаков использование такого классификатора в явном виде приводит к потере его ковариантности. Кроме того, в работе рассмотрен метод оценивания параметров логистической регрессии на основе метода максимального правдоподобия. Этот метод уже долгое время является одним из лучших для решения задач подобного вида. Это обусловлено рядом причин: актуальность и возможность применения в различных областях, а также возможность реализации метода на современных производительных компьютерах. Недостатком метода является его трудоёмкость.

В результате исследования определены, проанализированы и реализованы методы, которые дают возможность оценить вероятность заболевания пациента с заданными параметрами. Полученные данные позволят более точно оценивать состояние здоровья в условиях постоянно меняющихся диагностических параметров.

***Ключевые слова*:** классификация; оценка вероятности; логистическая регрессия; байесовский классификатор; метод оценки шансов; метод максимального правдоподобия.

Введение

Методы диагностики [1] в медицине играют важнейшую роль. Точность диагноза и быстрота, с которой его можно поставить, зависят от очень многих факторов: состояния больного, имеющихся данных о симптомах и признаках заболевания, результатах лабораторных анализов, но самое главное – от квалификации самого врача. Точно поставленный диагноз в кратчайшие сроки – увеличенный шанс на излечение больного. Исходя из всех этих соображений вполне естественно попытаться определить условия, при которых диагноз может быть поставлен максимально быстро и точно.

В течение многих веков врачи с переменным успехом предпринимали попытки решить эту задачу. Однако в последние годы благодаря применению современных методов лечения и диагностики, основанных на новейших достижениях науки и техники, возможности получения успешных результатов значительно возросли. Поэтому важно найти точные методы [2-3] описания,

исследования, оценки и контроля процесса постановки диагноза, что делает актуальной задачу определения вероятности заболевания на основании существующих данных о состоянии пациента.

Если исследование связано с большим числом взаимозависимых факторов, обнаруживающих значительную естественную изменчивость, то для достаточно эффективного описания сложной схемы их влияния существует лишь один способ – использование соответствующего статистического метода. Если существует необходимость определить вероятность попадания в один из двух классов заболевания, одним и наиболее простых и эффективных методов является бинарный классификатор. Качество классификации можно оценивать по каждой входной переменной отдельно. Если число факторов или число категорий данных очень велико – необходимо использовать вычислительные мощности компьютера [5], чтобы искомые результаты можно было получить за достаточно короткое время, что позволит сократить вероятность ошибки при постановке диагноза, а также сделать это максимально быстро и оперативно.

Таким образом, целью исследования является определения вероятности заболевания пациента [6-7] с заданными диагностическими характеристиками на основе методов Data Mining, что позволит повысить точность постановки диагноза.

1. Постановка задачи

На состояние здоровья каждого отдельного человека влияет целый ряд факторов, таких, как возраст, пол, пережитые болезни, место жительства, температура, состояние крови и т.д. Задача исследования состоит в определении и анализе методов, которые позволяют оценить вероятность заболевания [8] пациента с заданными диагностическими характеристиками.

Эту задачу относят к задачам классификации «с учителем», когда испытуемая система обучается с помощью примеров «стимул-реакция». Требуется найти зависимость, которая покажет, какие пациенты относятся к классу «Здоровые», а какие – к классу «Больные». Для такой задачи рационально использование логистической регрессии, широко используемой для нахождения вероятностей некоторого события при заданных характеристиках [10].

В статье предложены методы для оценивания параметров логистической регрессии для разных условий. В случае одной политомической входной переменной с минимальным числом категорий – метод оценки шансов и вероятностей. В данном случае качество классификации можно оценивать для каждой входной переменной отдельно: оценка не зависит от связанности входных переменных, что позволяет не проводить проверку коррелированности и предварительный отбор значимых переменных [10].

Для нескольких объясняющих переменных предлагается использовать байесовский классификатор, который позволит при отсутствии корреляции между признаками отнести конкретных индивидов рассматриваемой популяции к некоторому классу по состоянию здоровья. При наличии корреляции факторных признаков и сложных зависимостей между входными переменными предлагается использовать метод максимального правдоподобия. В результате анализа будет получен готовый математический аппарат, позволяющий на практике получить значения вероятностей заболеваний в условиях различных исходных данных.

2. Оценивание параметров логистической регрессии на основе метода оценки шансов и вероятностей

Рассмотрим некоторую выборку пациентов на основании данных из источника [12]. О каждом больном известна информация касающаяся состояния здоровья. Объясняющей переменной в данном случае является результат электрокардиограммы (ЭКГ) в состоянии покоя. Данная переменная является политомической. Каждый пациент может принадлежать по результатам ЭКГ к трем классам – «Normal», «Нур» и «Abnormal». Кроме того, исследуются два события: пациент болен (y=1) и здоров (y=0).

Необходимо оценить параметры логистического уравнения для данной задачи и определить, с какой вероятностью пациент будет принадлежать к классу «Болен», т.е. оценить его состояние здоровья.

Вероятность того, что выходная переменная $y=1$ для заданного значения объясняющей переменной x, будет $P(y=1|x)=\rho(x)$, а вероятность того, что $y=0$ при заданном значении x, будет равна $P(y=0|x)=1-\rho(x)$.

Условное среднее для логистической регрессии в данном случае определяется как в формуле (1):

$$\rho(x)=\frac{e^{g(x)}}{1+e^{g(x)}}, \quad (1)$$

где $g(x)=\beta_0+\beta_1 C_1+\beta_2 C_2$;

C_1, C_2 – переменные для квантования значений в трех интервалах;

x – объясняющая переменная;

$\beta_0, \beta_1, \beta_2$ – искомые параметры,

$\rho(x)$ – вероятность события.

Функция определена на бесконечном интервале и принимает значения в диапазоне $[0,1]$. Требуется найти наилучшие оценки параметров $\beta_0, \beta_1, \beta_2$. Упорядочим информацию касающуюся пациентов на основе данных [12] в виде таблицы (табл. 1).

Таблица 1

Данные о пациентах по состоянию ЭКГ

Исход	Normal	Нур	Abnormal	Всего
y=0	96	68	1	165
y=1	56	79	3	138
Всего	152	147	4	303

В табл. 1 в строке с классом «Normal» переменные квантования будут равны: $C_1=C_2=0$. В строке с классом «Нур» $C_1=1, C_2=0$. В строке с классом «Abnormal» $C_1=C_2=1$.

Шансы оказаться пациентом с больным сердцем для всех категорий состояний электрокардиограммы оценивается по формулам (2) – (4):

$$Ch_{y=1,C_1}=\frac{56}{96}\approx 0,58, \quad (2)$$

$$Ch_{y=1,C_2} = \frac{79}{68} \approx 1{,}16, \quad (3)$$

$$Ch_{y=1,C_3} = 3, \quad (4)$$

Отношение шансов для категорий «Нур» к категории «Normal» оценивается по формуле (5):

$$OR\left(\frac{C_2}{C_1}\right) = \frac{Ch_{y=1,C_2}}{Ch_{y=1,C_1}} = 2{,}01. \quad (5)$$

Отношение шансов для категорий «Abnormal» к категории «Normal» оценивается по формуле (6):

$$OR\left(\frac{C_3}{C_1}\right) = \frac{Ch_{y=1,C_3}}{Ch_{y=1,C_1}} = 5{,}17. \quad (6)$$

Экспериментальную вероятность заболевания для категории «Normal» можно найти (7), поделив число положительных исходов на общее количество исходов

$$c_{exp} = 56/152 = 0{,}37. \quad (7)$$

Отсюда коэффициент β_0 может быть найден как (8)

$$\beta_0 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) = -0{,}532. \quad (8)$$

Для категории «Нур» экспериментальную вероятность заболевания можно оценить по формуле (9):

$$c_{exp} = 79/147 = 0{,}54. \quad (9)$$

Отсюда коэффициент β_1 может быть найден как (10):

$$\beta_1 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) - \beta_0 = 0{,}692. \quad (10)$$

Для категории «Abnormal» экспериментальную вероятность заболевания можно оценить по формуле (11):

$$c_{exp} = 3/4 = 0{,}75. \quad (11)$$

Отсюда коэффициент β_2 может быть найден как (12)

$$\beta_2 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) - \beta_0 - \beta_1 = 0{,}939. \quad (12)$$

Вероятность того, что выходная переменная y будет равна единице (т.е. пациент будет болен) для категории «Normal» рассчитывается по формуле (13):

$$P(y=1\,|\,x) = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-0{,}532}}{1+e^{-0{,}532}} \approx 0{,}37. \quad (13)$$

Вероятность того, что выходная переменная $y=1$ для категории «Нур», рассчитывается по формуле (14):

$$P(y=1|x)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}=\frac{e^{-0,532+0,692}}{1+e^{-0,532+0,692}}\approx 0,54. \quad (14)$$

Вероятность того, что выходная переменная $y=1$ для категории «Abnormal» рассчитывается по формуле (15):

$$P(y=1|x)=\frac{e^{\beta_0+\beta_1+\beta_2}}{1+e^{\beta_0+\beta_1+\beta_2}}=\frac{e^{-0,532+0,692+0,939}}{1+e^{-0,532+0,692+0,939}}\approx 0,75. \quad (15)$$

Можно сделать вывод о том, что если результат ЭКГ «Abnormal», то вероятность заболевания выше всего, если «Нур» – то меньше, а вероятность оказаться здоровым выше всего – если результат «Normal».

3. Оценка вероятности заболевания с использованием байесовского классификатора

Рассмотрим выборку из 30 пациентов входными переменными, заданными в номинальной шкале (табл. 2) на основании данных из источника [12]. Для анализа используем следующие признаки: возраст (в годах), сахар в крови, пол пациента, результат ЭКГ. По данным табл. 2 были рассчитаны коэффициенты парной корреляции, значения которых находятся в интервале [-0.303; 0.078], что свидетельствует о низкой корреляционной зависимости между входными переменными.

Таблица 2

Данные о пациентах по выбранным характеристикам

n/n	Возраст	A1	A2	A3	A4
1	60-69	Нет	Муж.	Нур	Да
2	40-49	Нет	Муж.	Normal	Нет
3	60-69	Нет	Муж.	Нур	Нет
4	60-69	Нет	Муж.	Normal	Да
5	50-59	Да	Муж.	Нур	Да
6	50-59	Да	Жен.	Normal	Нет
7	50-59	Нет	Муж.	Normal	Да
8	50-59	Нет	Муж.	Normal	Да
9	60-69	Да	Муж.	Normal	Нет
10	60-69	Нет	Муж.	Нур	Да
11	60-69	Нет	Муж.	Нур	Да
12	30-39	Нет	Муж.	Normal	Нет
13	40-49	Нет	Жен.	Normal	Нет
14	50-59	Нет	Муж.	Normal	Нет
15	60-69	Нет	Жен.	Normal	Да
16	50-59	Нет	Жен.	Нур	Нет
17	50-59	Нет	Муж.	Normal	Нет
18	50-59	Нет	Жен.	Normal	Нет
19	50-59	Да	Муж.	Нур	Да
20	40-49	Нет	Муж.	Нет	Нет
21	50-59	Нет	Жен.	Нет	Нет
22	50-59	Нет	Жен.	Normal	Нет
23	60-69	Нет	Жен.	Normal	Нет

Окончание табл. 2

n/n	Возраст	A1	A2	A3	A4
24	40-49	Нет	Муж.	Normal	Нет
25	40-49	Нет	Муж.	Нур	Да
26	60-69	Нет	Жен.	Normal	Нет
27	60-69	Да	Муж.	Нур	Да
28	60-69	Нет	Муж.	Normal	Да
29	50-59	Нет	Муж.	Normal	Нет
30	40-49	Нет	Муж.	Нур	Нет

*Примечание. **A1** – сахар в крови менее 120 ед.; **A2** – пол пациента; **A3** – результат ЭКГ; **A4** – состояние болезни.*

Таким образом, в данном случае можно использовать байесовский классификатор, применение которого для этого случая подробно рассмотрено в работе [10].

Обозначим как С1 класс «Больные», для которых состояние болезни присутствует (значение результирующей переменной – «да»). Через С2 можно обозначить класс пациентов «Здоровые», который не имеют признаков болезни (значение результирующей переменной – «нет»).

Применение байесовского классификатора не дает возможности получить вид статистической зависимости на основе обучающей выборки, однако позволяет определить вероятность того, что пациент с заданными характеристиками попадет в тот или иной класс. Например, определим, что пациент возраста от 50 до 59 лет, с сахаром в крови менее 120 единиц, мужчина и с результатом ЭКГ «Нур» попадет в класс «Больные».

Необходимо максимизировать произведение вероятностей $P(X \mid C_k)P(C_k)$ для $k = 2$, так как в данной задаче всего два класса. Априорная вероятность появления класса C_1 вычисляется по формуле (16):

$$P(C_1) = \frac{12}{30} = 0,4. \quad (16)$$

Априорная вероятность появление класса C_2 вычисляется по формуле (17):

$$P(C_2) = \frac{18}{30} = 0,6. \quad (17)$$

Всего наблюдаемых примеров 30, 18 из них – «Здоровые», 12 – «Больные». Условные вероятности для определения $P(X \mid C_k)$ рассчитаны в табл. 3. Рассчитаем обобщенные вероятности $P(X \mid C_k)$ для событий формулы (18-19):

$$P(X \mid C_1) = 0,33 \cdot 0,25 \cdot 0,92 \cdot 0,58 = 0,044; \quad (18)$$

$$P(X \mid C_2) = 0,44 \cdot 0,11 \cdot 0,55 \cdot 0,17 = 0,0045. \quad (19)$$

Тогда вероятности $P(X \mid C_k)P(C_k)$ будут соответственно равны (20-21):

$$P(X \mid C_1)P(C_1) = 0,044 \cdot 0,6 = 0,0264; \quad (20)$$

$$P(X \mid C_2)P(C_2) = 0,0045 \cdot 0,4 = 0,0018. \quad (21)$$

Таблица 3

Условные вероятности для данных о пациентах

Описание вероятности	Расчет
$P(Возраст\,50-59\,|\,C_2)$	8/18=0,44
$P(Возраст\,50-59\,|\,C_1)$	4/12=0,33
$P\left(Сахар\,в\,крови<120\,|\,C_2\right)$	2/18=0,11
$P\left(Сахар\,в\,крови<120\,|\,C_1\right)$	3/12=0,25
$P(Мужчина\,|\,C_2)$	10/18=0,55
$P(Мужчина\,|\,C_1)$	11/12=0,92
$P(Нур\,|\,C_2)$	3/18=0,17
$P(Нур\,|\,C_1)$	7/12=0,58

Выбирается тот класс, вероятность для которого больше, т.е. рассматриваемый пациент относится к классу C_1 – «Больной».

Нормализация вероятностей может выглядеть следующим образом формулы (22-23):

$$P'(X\,|\,C_1)P\left(C_1\right)=\frac{0,0264}{0,0264+0,0018}=0,94\,; \tag{22}$$

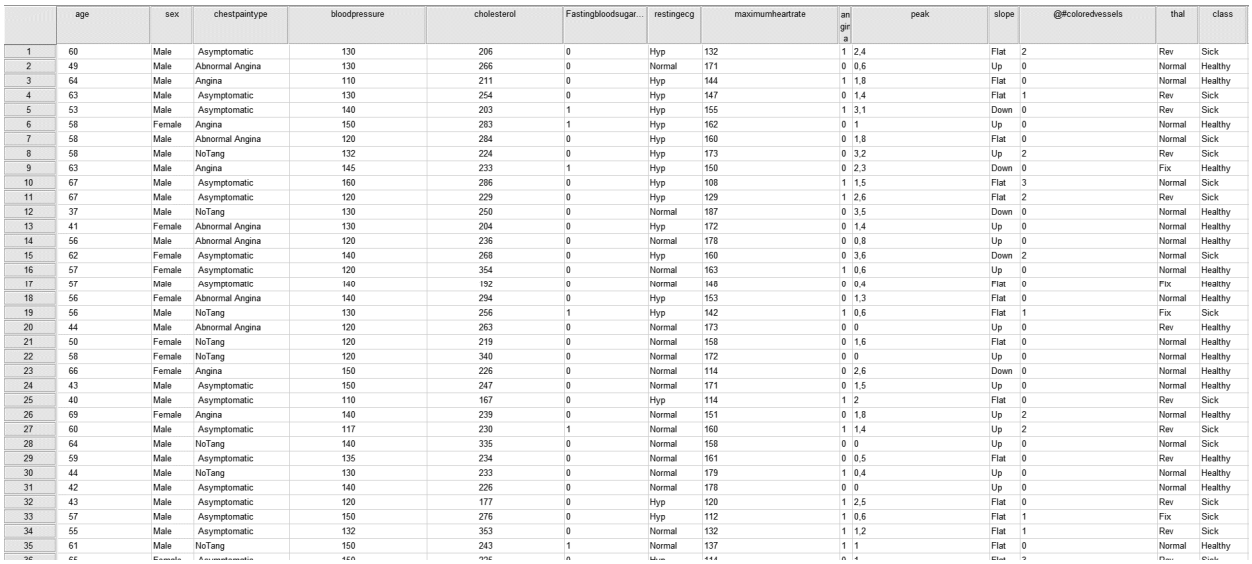
$$P'(X\,|\,C_2)P\left(C_2\right)=\frac{0,0018}{0,0018+0,0264}=0,06\,. \tag{23}$$

Таким образом, пациент с описанными характеристиками с вероятностью 0.94 окажется больным (попадет в класс «Больные»), а с вероятностью 0.06 окажется здоровым (попадет в класс «Здоровые»).

4. Оценивание параметров логистической регрессии на основе метода максимального правдоподобия

Рассмотрим выборку из 303 пациентов с входными характеристиками, показанными на рисунке на основании данных из источника [12]. Результирующий признак измеряется в дихотомической шкале, а факторные признаки – в метрических и других видах шкал. Необходимо определить вероятность болезни пациента с данным множеством характеристик.

Так как метод максимального правдоподобия (ММП) является достаточно ресурсоёмким, то для демонстрации будем использовать программное обеспечение (ПО) от IBM – SPSS Statistics. Данное ПО позволяет не только найти параметры логистической регрессии, но и оценить параметры модели и вероятности, также проанализировать качество модели.



	age	sex	chestpaintype	bloodpressure	cholesterol	Fastingbloodsugar...	restingecg	maximumheartrate	an gin a	peak	slope	@#coloredvessels	thal	class
1	60	Male	Asymptomatic	130	206	0	Hyp	132	1	2,4	Flat	2	Rev	Sick
2	49	Male	Abnormal Angina	130	266	0	Normal	171	0	0,6	Up	0	Normal	Healthy
3	64	Male	Angina	110	211	0	Hyp	144	1	1,8	Flat	0	Normal	Healthy
4	63	Male	Asymptomatic	130	254	0	Hyp	147	0	1,4	Flat	1	Rev	Sick
5	53	Male	Asymptomatic	140	203	1	Hyp	155	1	3,1	Down	0	Rev	Sick
6	58	Female	Angina	150	283	1	Hyp	162	0	1	Up	0	Normal	Healthy
7	58	Male	Abnormal Angina	120	284	0	Hyp	160	0	1,8	Flat	0	Normal	Sick
8	58	Male	NoTang	132	224	0	Hyp	173	0	3,2	Up	2	Rev	Sick
9	63	Male	Angina	145	233	1	Hyp	150	0	2,3	Down	0	Fix	Healthy
10	67	Male	Asymptomatic	160	286	0	Hyp	108	1	1,5	Flat	3	Normal	Sick
11	67	Male	Asymptomatic	120	229	0	Hyp	129	1	2,6	Flat	2	Rev	Sick
12	37	Male	NoTang	130	250	0	Normal	187	0	3,5	Down	0	Normal	Healthy
13	41	Female	Abnormal Angina	130	204	0	Hyp	172	0	1,4	Up	0	Normal	Healthy
14	56	Male	Abnormal Angina	120	236	0	Normal	178	0	0,8	Up	0	Normal	Healthy
15	62	Female	Asymptomatic	140	268	0	Hyp	160	0	3,6	Down	2	Normal	Sick
16	57	Female	Asymptomatic	120	354	0	Normal	163	1	0,6	Up	0	Normal	Healthy
17	57	Male	Asymptomatic	140	192	0	Normal	148	0	0,4	Flat	0	Fix	Healthy
18	56	Female	Abnormal Angina	140	294	0	Hyp	153	0	1,3	Flat	0	Normal	Healthy
19	56	Male	NoTang	130	256	1	Hyp	142	1	0,6	Flat	1	Fix	Sick
20	44	Male	Abnormal Angina	120	263	0	Normal	173	0	0	Up	0	Rev	Healthy
21	50	Female	NoTang	120	219	0	Normal	158	0	1,6	Flat	0	Normal	Healthy
22	58	Female	NoTang	120	340	0	Normal	172	0	0	Up	0	Normal	Healthy
23	66	Female	Angina	150	226	0	Normal	114	0	2,6	Down	0	Normal	Healthy
24	43	Male	Asymptomatic	150	247	0	Normal	171	0	1,5	Up	0	Normal	Healthy
25	40	Male	Asymptomatic	110	167	0	Hyp	114	1	2	Flat	0	Rev	Sick
26	69	Female	Angina	140	239	0	Normal	151	0	1,8	Up	2	Normal	Healthy
27	60	Male	Asymptomatic	117	230	1	Normal	160	1	1,4	Up	2	Rev	Sick
28	64	Male	NoTang	140	335	0	Normal	158	0	0	Up	0	Normal	Sick
29	59	Male	Asymptomatic	135	234	0	Normal	161	0	0,5	Flat	0	Rev	Healthy
30	44	Male	NoTang	130	233	0	Normal	179	1	0,4	Up	0	Normal	Healthy
31	42	Male	Asymptomatic	140	226	0	Normal	178	0	0	Up	0	Normal	Healthy
32	43	Male	Asymptomatic	120	177	0	Hyp	120	1	2,5	Flat	0	Rev	Sick
33	57	Male	Asymptomatic	150	276	0	Hyp	112	1	0,6	Flat	1	Fix	Sick
34	55	Male	Asymptomatic	132	353	0	Normal	132	1	1,2	Flat	1	Rev	Sick
35	61	Male	NoTang	150	243	1	Normal	137	1	1	Flat	0	Normal	Healthy

Фрагмент данных о выборке пациентов в файле SPSS

Самые значимые результаты видны в таблицах ниже. В табл. 4 представлены коэффициенты качества модели.

Таблица 4

Сводная таблица модели

Шаг	-2 Log правдоподобие	R-квадрат Кокса и Снелла	R-квадрат Нейджелкерка
1	348.461	0.210	0.281

Критерий -2 Log правдоподобия характеризует соответствие модели и исходных данных. Чем меньше этот показатель, тем адекватнее модель. R-квадрат Кокса и Снелла и R-квадрат Нейджелкерка устойчивее традиционных статистик согласия, которые используются в логите. По первой характеристике значение равное единице, достижимо. Во второй характеристике этот недостаток устранен. Данные критерии показывают долю влияния всех факторных признаков на дисперсию зависимой переменной. Более подробную информацию можно взять из источника [13]. В табл. 5 представлены значения критерия Хи-квадрат.

Таблица 5

Универсальный критерий коэффициентов модели

Шаг		Хи-квадрат	Степени свободы	Значимость
1	Шаг	71,586	3	,000
	Блок	71,586	3	,000
	Модель	71,586	3	,000

В табл. 6-7 представлен критерий Хосмера-Лемешова. В нашем случае часть дисперсии составляет 0.7%. Это свидетельствует о высокой степени согласованности модели.

Критерий Хосмера-Лемешова – показывает оценку согласия между частотами в выборке и моделью [14]. Он показывает, присутствует ли «мусор», который приводит к снижение качества модели, в модели.

Таблица 6

Критерий Хосмера-Лемешова

Шаг	Хи-квадрат	Степени свободы	Значимость
1	9.800	2	0.007

Таблица 7

Таблица сопряженности для проверки согласия Хосмера-Лемешова

		Болезнь = да		Болезнь = нет		Всего
		Наблюдаемые	Ожидаемые	Наблюдаемые	Ожидаемые	
Шаг 1	1	23	22,839	0	0,161	23
	2	27	29,294	3	0,706	30
	3	29	28,613	1	10,387	30
	4	27	26,486	3	3,514	30
	5	27	23,015	3	6,985	30
	6	16	17,586	14	12,414	30
	7	10	11,150	20	18,850	30
	8	5	4,997	24	24,003	29
	9	0	1,847	30	28,153	30
	10	1	0,495	40	40,505	41

В табл. 8 приведены проценты, отображающие разные уровни классификации модели. Получены достаточно высокие показатели, т.е. 92.2% случаев удалось классифицировать верно.

Таблица 8

Таблица классификации

Наблюдаемые			Предсказанные		
			Class		Процент правильных
			Healthy	Sick	
Шаг 1	Class	Healthy	147	18	93,1
		Sick	23	115	91,3
	Общая процентная доля				92,2

В табл. 9 приведены параметры уравнения логистической регрессии.

Таблица 9

Переменные уравнения регрессии

Влияющая переменная	Коэффициент уравнения регрессии β	Среднеквадратическая ошибка	Статистика Вальда	Уровень значимости
A – sex(1)	-1,464	0,490	8,932	0,003
B – chestpaintype			30,864	0,000
B1 – chestpaintype(1)	2,286	0,444	26,474	0,000
B2 – chestpaintype(2)	0,971	0,590	2,705	0,100

Окончание табл. 9

Переменные уравнения регрессии

Влияющая переменная	Коэффициент уравнения регрессии β	Среднеквадратическая ошибка	Статистика Вальда	Уровень значимости
B3 - chestpaintype(3)	0,170	0,652	0,068	0,794
C - angina(1)	-0,763	0,380	4,044	0,044
D – slope			24,588	0,000
D1 - slope(1)	1,724	0,700	6,068	0,014
D2 - slope(2)	2,018	0,415	23,700	0,000
E - @#coloredvessels			36,481	0,000
E1 - @#coloredvessels(1)	-1,763	0,505	12,204	0,000
E2 - @#coloredvessels(2)	0,495	0,547	0,818	0,366
E3 - @#coloredvessels(3)	1,498	0,793	3,566	0,059
F – thal			14,056	0,001
F1 - thal(1)	-1,492	0,733	4,137	0,042
F2 - thal(2)	-1,452	0,411	12,472	0,000

Остальные переменные были исключены из формулы из-за избыточности данных.

На основе данной таблицы можно определить наиболее значимые факторы, по которым можно получить наименьшие ошибки с высокой долей вероятности. Общий вид уравнения регрессии для пациента будет иметь вид аналогичный формуле (24):

$$g(x) = -1,464 \cdot \mathrm{A} + 2,286 \cdot \mathrm{B1} + 0,971 \cdot \mathrm{B2} + \\ +0,170 \cdot \mathrm{B3} - 0,763 \cdot \mathrm{C} + 1,724 \cdot \mathrm{D1} + 2,018 \cdot \mathrm{D2} - \\ -1,763 \cdot \mathrm{E1} + 0,495 \cdot \mathrm{E2} + 1,498 \cdot \mathrm{E3} - 1,492 \cdot \mathrm{F1} - 1,452 \cdot \mathrm{F2} \quad (24)$$

Тогда для пациента мужского пола со вторым типом боли в груди, который не болел ангиной, с уклоном первого типа, с сосудами третьего типа, а также thal'ом первого типа будет справедливо (25):

$$g(x) = -1,464 + 0,971 + 1,724 + 1,498 - 1,492 = 1,237. \quad (25)$$

Вероятность того, что такой пациент окажется здоровым, вычисляется по формуле (26):

$$\rho(x) = \frac{e^{g(x)}}{\left(1 + e^{g(x)}\right)} \approx 0,77. \quad (26)$$

При этом, как видно из табл. 9, по статистике Вальда наиболее значимыми являются такие факторы: chestpaintype (значение 30,864), slope (значение 24,588), coloredvessels (значение 36,481).

Тест Вальда — статистический тест, используемый для проверки ограничений на параметры статистических моделей, оценённых на основе

выборочных данных. Является самым приемлемым из трёх базовых тестов проверки ограничений, таких, как тест отношения правдоподобия и тест множителей Лагранжа. Тест является асимптотическим, т.е. для достоверности выводов требуется достаточно большой объём выборки. Кроме того доверительный интервал (ДИ) теста представляет собой замкнутую форму. Чем выше значение статистики, тем лучше.

Значимость факторов подтверждаем с помощью соответствующего уровня значимости. Она определяется как p-уровень, который рассчитывается в ходе теста. Чем меньше этот уровень, тем лучше.

На основе данных на рис. 1 были рассчитаны вероятности попаданий в группу «Здоров» для всех данных.

В столбцах «Ожидаемые» и «Группа» можно увидеть вероятности попадания в группу «Здоров» или «Болен».

На основе результатов можно сказать, что модель адекватно описывает данную совокупность.

Выводы

Таким образом, в данной работе определены, проанализированы и реализованы методы, которые позволяют оценить вероятность заболевания пациента с заданными диагностическими характеристиками.

Показано, в каких случаях целесообразно применение тех или иных методов для определения вероятности и оценивания параметров моделей. Данные модели не являются статичными. Расчет параметров можно проводить каждый раз при изменении количества данных о пациентах, а использование программного инструментария SPSS позволит осуществлять расчеты достаточно оперативно. Полученные данные позволят более точно оценивать состояние здоровья в условиях постоянно меняющихся диагностических параметров.

Список литературы (References)

1. P. Baldi and S. Brunak, Bioinformatics: The Machine Learning Approach (2nd ed.) [Text] / MIT Press, 2001.

2. M. W. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval [Text] / Springer, 2003.

3. J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability [Text] / Berkeley, University of California Press, 1967 – pg. 281-297.

4. M. Deshpande, Automated approaches for classifying structures. In Proc. 2002Workshop onDataMining in Bioinformatics (BIOKDD'02) [Text] / M. Deshpande, M. Kuramochi, G. Karypis Edmonton. Canada, 2002 – pg. 11–18.

5. W. Frakes, Information Retrieval: Data Structures and Algorithms [Text] / W. Frakes, R. Baeza-Yates. Prentice Hall, 1992.

6. International Agency for Research on Cancer [Electronic resource] / Lyon, France, 2013, Access Mode: http://globocan.iarc.fr.

7. Cancer control: early detection. WHO Guide for effective programmes. [Electronic resource] / Geneva: World Health Organization; 2007, Access Mode: http://apps.who.int/iris/bitstream/10665/43743/1/9241547338_eng.pdf.

8. Rubin G, The expanding role of primary care in cancer control [Text] / Rubin G, Berendsen A, Crawford SM, Dommett R, Earle C. Lancet Oncol, 2015 – pg. 31–72.

9. Bowers, N.L., Actuarial mathematic [Text] / Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., Nesbitt, C.J. Schaumburg, Illinois, USA by Society Of Actuaries, (1997) – pg 621.

10. Norman T. J. Bailey, The mathematical approach to biology and medicine norman [Text] / Norman T. J. Bailey. Wiley, 1967 – pg. 296.

11. Tom I.Je. Tehnologija analiza medicinskih dannyh statisticheskimi i nejrosetevymi metodami [Text] / I.Je. Tom, O.V. Krasko, N.A. Novoselova, M.P. Potapnev, T.A. Uglova // Iskusstvennyj intellekt. – 2004. – №2. – pg. 372-376.

12. Supplemental Excel Data Sets [Electronic resource] / Access Mode: http://mercury.webster.edu/aleshunas/Data%20Sets/Supplemental%20Excel%20Data%20Sets.htm.

13. Cox D.R. Analysis of Binary Data [Text] / D.R.Cox, E.J. Snell.– Chapman and Hall / CRC, 1989. – pg. 240.

14. Hosmer-Lemeshow Test [Electronic resource] / Access Mode: http://www.real-statistics.com/logistic-regression/hosmer-lemeshow-test/.

Поступила в редакцию 19.03.2019, рассмотрена на редколлегии 20.03.2019

Визначення ймовірності захворювання хворобами серця на основі методів Data Mining

У сучасному світі фахівці різних областей шукають способи і методи лікування і діагностування захворювань. Рішення проблеми, яка полягає в обмежених можливостях своєчасної діагностики, лежить в області методів Data Mining.

Першим із методів, що розглядаються в роботі, є метод оцінювання параметрів логістичної регресії на основі методу оцінки шансів і вірогідності. Метод найвигідніше використовувати для вибірок із малою кількістю параметрів. На вибірці з великою кількістю параметрів цей метод перестає бути актуальним і втрачає свою точність. Другим методом є метод оцінки ймовірності захворювання з використанням байєсівського класифікатора. Цей метод вигідніше використовувати на вибірках із великою кількістю параметрів, тому що метод не втрачає своєї точності при збільшенні кількості змінних, однак, незважаючи на це кількість ознак має бути постійним. У випадку з перемінною кількістю ознак використання такого класифікатора в явному вигляді призводить до втрати його коваріантности. Крім того, в роботі розглядається метод оцінювання параметрів логістичної регресії на основі методу максимальної правдоподібності. Цей метод вже довгий час є одним із кращих для вирішення завдань подібного виду. Це обумовлено рядом причин: актуальність і можливість застосування в різних областях, а також можливість реалізації методу на сучасних продуктивних комп'ютерах. Недоліком методу є його трудомісткість.

У результаті дослідження визначено, проаналізовано і реалізовано методи, які дозволяють оцінити ймовірність захворювання пацієнта з заданими параметрами. Отримані дані дозволять більш точно оцінювати стан здоров'я в умовах постійно мінливих діагностичних параметрів.

Ключові слова: класифікація; оцінка ймовірності; логістична регресія; байєсівський класифікатор; метод оцінки шансів; метод максимальної правдоподібності.

Determining the probability of heart disease based on Data Mining methods

In the modern world, when people suffer from various diseases, many experts are looking for ways and methods to treat and diagnose them. The solution to the problem, which lies in the limited possibilities of timely diagnosis, lies in the field of Data Mining methods.

The first of the methods considered in the paper is the method of estimating logistic regression parameters based on the method of estimating odds and probabilities. The method is most advantageous to use for samples with a small number of parameters. On a sample with a large number of parameters, this method ceases to be relevant and loses its accuracy. The second method considered is a method for estimating the probability of a disease using a Bayesian classifier. It is more profitable to use this method on samples with a large number of parameters, since the method does not lose its accuracy with an increase in the number of variables, however, despite this number of signs should be constant. In the case of a variable number of attributes, the use of such a classifier in an explicit form leads to the loss of its covariance. The paper also discusses a method for estimating logistic regression parameters based on the maximum likelihood method. This method has long been one of the best for solving problems of this type. This is due to several reasons: the relevance and the possibility of application in various fields, as well as the possibility of implementing the method on modern productive computers. The disadvantage of the method is its complexity.

As a result of the study, methods were determined, analyzed and implemented that allow to estimate the probability of the patient's disease with the given parameters. The obtained data will allow to more accurately assess the state of health in the conditions of constantly changing diagnostic parameters.

Keywords: classification; probability estimation; logistic regression; Bayes classifier; odds ration method; maximum likelihood method.

Сведения об авторах:

Базилевич Ксения Алексеевна − доцент кафедры математического моделирования и искусственного интеллекта, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, k.bazilevych@khai.edu, 0000-0001-5332-9545.

Меняйлов Евгений Сергеевич − старший преподаватель кафедры математического моделирования и искусственного интеллекта, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, j.menyailov@khai.edu, 0000-0002-9440-8378.

Горанина Сергей Игоревич − студент кафедры математического моделирования и искусственного интеллекта, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, sgoranin@gmail.com, 0000-0001-8988-3935.

Федулов Кирилл Андреевич − студент кафедры математического моделирования и искусственного интеллекта, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, fedulov.kirill172@gmail.com, 0000-0001-9619-0299.