



М.А. ИВАНЧУК, И.В. МАЛЫК

УДК [519.237.8+519.244.3]:616.079.5

**СРАВНЕНИЕ МЕТОДОВ
РАСПРЕДЕЛЕНИЯ НАБЛЮДЕНИЙ
НА КЛАССЫ ПРИ ПРОГНОЗИРОВАНИИ
НАЛИЧИЯ ОСЛОЖНЕНИЙ
У ТЯЖЕЛОБОЛЬНЫХ**

Аннотация. Проведено сравнение кластерного анализа, вальдовского последовательного анализа, а также собственного метода поворота системы координат для распределения на классы больных с острым деструктивным панкреатитом и пациентов с острыми сочетанными радиационными поражениями для ранней диагностики наличия у них осложнений. В первом случае наиболее эффективным оказался вальдовский анализ, во втором случае для практического использования предложены кластерный анализ методом k -средних и метод поворота координат.

Ключевые слова: кластерный анализ, информационная мера Кульбака, последовательный анализ Вальда.

ВВЕДЕНИЕ

Методы распределения признаков на классы в последнее время все чаще используются в медицине при раннем прогнозировании тяжести заболевания [1–4]. В настоящей статье на примере двух клинических групп рассмотрено применение кластерного анализа, информационной меры Кульбака с последующим вальдовским анализом, а также собственного метода поворота системы координат для прогнозирования наличия осложнений у тяжелобольных.

1. ПОСТАНОВКА ЗАДАЧИ

Для сравнения рассматриваемых методик используем две выборки.

1. *Больные с острым деструктивным панкреатитом (ОДП), у которых необходимо прогнозировать наличие генных модификаций.* Наличие исследуемого генетического дефекта у больных с ОДП может существенно влиять на характер поражения поджелудочной железы и прилегающих к ней тканей и, как следствие, негативно отражаться на конечных результатах лечения. Раннее прогнозирование наличия генных модификаций, исходя из результатов анализов, полученных на этапе поступления больного в клинику, позволяет врачу адекватно выбрать тактику лечения таких больных [5].

2. *Больные с острыми сочетанными радиационными поражениями (ОСРП), степень тяжести которых необходимо прогнозировать.* В настоящее время большое значение имеет теоретический анализ количественной связи дозы радиации с тяжестью облучений. На данное время оценен вклад костномозгового, желудочно-кишечного и нервного синдромов в возможность гибели при высоких дозах облучений [6]. Определена доля лучевой нагрузки при облучении людей в условиях ядерного взрыва, построены графики кривых летальности, рассмотрен дозиметрический критерий для острой лучевой болезни, найдена эмпиричес-

© М.А. Иванчук, И.В. Малык, 2015

кая зависимость выживания от дозы облучения для лабораторных животных [7–9]. Цель нашего исследования — раннее прогнозирование летального исхода и степени тяжести ОСРП для правильного и своевременного распределения больных с различными формами заболевания во время аварийных ситуаций.

Будем считать, что нулевая гипотеза H_0 означает отсутствие осложнений у больных. Тогда альтернативная ей гипотеза H_1 означает их наличие. При этом ошибка первого рода α будет означать риск гипердиагностики (т.е. отнесение больного с более легкой формой заболевания в группу тяжелобольных), а ошибка второго рода β — риск гиподиагностики (т.е. отнесение тяжелобольного в группу с легким течением заболевания) [10]. Безусловно, при такой формулировке гипотез для врача и пациента ошибки второго рода являются более важными, чем ошибки первого рода.

Для выбора лучшего критерия прогнозирования будем сравнивать их мощность Se и чувствительность Sp , вычисленные по формулам [10]

$$Se = \frac{a}{a+c}, \quad (1)$$

$$Sp = \frac{d}{b+d}, \quad (2)$$

где a — количество истинно положительных результатов, b — количество ложноположительных результатов, c — количество ложноотрицательных результатов, d — количество истинно отрицательных результатов.

1.1. Наличие генных модификаций у больных с ОДП. В качестве маркеров наличия генных модификаций у n больных ($n=88$) с ОДП рассматриваем m показателей ($m=21$). Результаты общего анализа крови: количество лейкоцитов, глюкозы, кальция, прокальцитонина, липазы, трипсина; гематокрит; концентрация амилазы в моче; температура тела; индексы тяжести состояния больных: BMI, APACHE II, SAPS, MODS, SOFA, Baltazar; временные показатели: время до госпитализации, сроки формирования осложнений, количество оперативных вмешательств, количество койко-дней, количество повторных госпитализаций [5]. Больные были разделены на две группы: обучающую (20 больных с наличием генных мутаций и 50 больных без них) и контрольную (8 больных с генными мутациями и 10 больных без них).

1.2. Степень тяжести ОСРП. Расчеты проводились на основе данных из историй болезни особ ($n=35$), пострадавших во время аварии на Чернобыльской АЭС [11]. Выделим три группы пострадавших в зависимости от тяжести их состояния и сроков летального исхода: 1-я группа — пострадавшие в очень тяжелом состоянии, умерли на 14–21-й день после аварии; 2-я группа — тяжелое состояние пострадавших обусловило летальность, начиная с четвертой недели после облучения; 3-я группа — тяжесть облучения не привела к смерти пострадавших.

В качестве маркеров для прогнозирования тяжести ОСРП будем учитывать m ($m=2$) следующих признаков: дозу общего облучения R и характеристику площади поражения и тяжести радиационного поражения кожи B : $B = A \times S$,

где $S = \sum_{j=1}^4 S_j$, S_j — процент площади поверхности тела с поражением j -й степени ($j = \overline{1, 4}$).

Коэффициент A будем вычислять по формуле [12]

$$A = \begin{cases} 0,5, & \text{если } S_1 + S_2 < 30 \% ; S_3 + S_4 < 15 \% , \\ 1, & \text{если } S_1 + S_2 \geq 30 \% ; S_3 + S_4 < 15 \% , \\ 1,5, & \text{если } S_1 + S_2 < 30 \% ; 15 \% \leq S_3 + S_4 < 100 \% , \\ 1,5, & \text{если } S_1 + S_2 \geq 30 \% ; 15 \% \leq S_3 + S_4 < 30 \% , \\ 2, & \text{если } S_1 + S_2 \geq 30 \% ; S_3 + S_4 \geq 30 \% . \end{cases}$$

2. ДИСПЕРСИОННЫЙ АНАЛИЗ

Проверим, являются ли отобранные признаки маркерами наличия осложнений. Воспользуемся дисперсионным анализом Фишера–Снедекора [13]. Для этого найдем экспериментальные значения критерия Фишера по формуле

$$F = \sigma_f^2 / \sigma_r^2, \quad (3)$$

где $\sigma_f^2 = n\sigma_{\bar{x}}^2$ — факторная дисперсия, $\sigma_r^2 = (\sigma_1^2 + \sigma_2^2 + \dots + \sigma_k^2) / k$ — остаточная дисперсия, и сравним найденное значение с критическим.

Оценку силы влияния фактора будем проводить по методу Снедекора:

$$h_x^2 = \frac{\sigma_f^2 - \sigma_r^2}{\sigma_f^2 - \sigma_r^2 + \bar{n} \cdot \sigma_r^2}, \quad (4)$$

где $\bar{n} = \frac{1}{k-1} \left(N - \frac{1}{N} \sum_{i=1}^k q_i^2 \right)$, $N = \sum_{i=1}^k q_i$, k — число градаций фактора, q_i — число вариантов в отдельных градациях фактора.

2.1. Наличие генных модификаций у больных с ОДП. На уровне значимости $p < 0,01$ такие факторы, как время до госпитализации, количество в крови глюкозы, кальция, прокальцитонина, липазы, трипсина; гематокрит; концентрация амилазы в моче; температура тела, сроки формирования осложнений, количество оперативных вмешательств, количество койко-дней, индексы APACHE II, SAPS, MODS, SOFA, Baltazar достоверно отличаются у больных, имеющих генные модификации, от больных, не имеющих их.

2.2. Степень тяжести ОСРП. На уровне значимости $p < 0,01$ оба фактора (доза общего облучения и характеристика радиационных поражений кожи) достоверно влияют на степень тяжести ОСРП. Сила влияния дозы составляет 43 %, а радиационных поражений кожи — 56 %.

Значит, отобранные нами признаки действительно можно использовать в качестве маркеров наличия осложнений.

3. ОЦЕНКА ИНФОРМАТИВНОСТИ ПО КУЛЬБАКУ, ПОСЛЕДОВАТЕЛЬНЫЙ АНАЛИЗ ВАЛЬДА

Воспользуемся функционалом Кульбака для вычисления информативности признаков и нахождения соответствующих диагностических коэффициентов [14–16].

Диагностические коэффициенты (ДК) вычислим по формуле

$$ДК(x_{ij}) = 10 \cdot \lg \frac{P(x_{ij} / G_+)}{P(x_{ij} / G_-)}, \quad (5)$$

где G_+ — наличие осложнений у больного; G_- — отсутствие осложнений; x_{ij} — j -й диапазон прогностического критерия x_i , $i = 1, m$; $P(x_{ij} / G_+)$ — условная вероятность попадания больного с осложнениями в j -й диапазон i -го прогностического критерия; $P(x_{ij} / G_-)$ — условная вероятность попадания больного без осложнений в j -й диапазон i -го прогностического критерия.

Для определения вклада данного прогностического критерия в приближение к верному диагностическому порогу вычислим информативность критериев [14] по формуле

$$J(x_i / G_+, x_i / G_-) = \sum_{j=1}^{n_i} 10 \cdot \lg \frac{P(x_{ij} / G_+)}{P(x_{ij} / G_-)} \times 0,5 [P(x_{ij} / G_+) - P(x_{ij} / G_-)], \quad (6)$$

где x_i — прогностический критерий, n_i — количество диапазонов прогностического критерия x_i .

Будем распределять больных по группам, используя вальдовский последовательный метод, который состоит в последовательном накоплении информации до момента достижения прогностического порога. Этот подход является отличительной чертой последовательного статистического анализа от классического анализа и требует в среднем в два раза меньше информации для принятия решения с заданным уровнем надежности [14].

Пороги для сумм диагностических коэффициентов при выборе гипотезы, касающейся наличия осложнений у больных, будем определять соотношениями

$$\sum_{i=1}^m ДК_i \geq 10 \cdot \lg \frac{\alpha}{1-\beta} \quad (7)$$

для нижней границы и

$$\sum_{i=1}^m ДК_i \leq 10 \cdot \lg \frac{1-\alpha}{\beta} \quad (8)$$

для верхней границы, где $ДК_i$ — диагностический коэффициент i -го признака, α и β — ошибки первого и второго рода соответственно.

3.1. Наличие генных модификаций у больных с ОДП. Диагностические коэффициенты и информативность будем вычислять по формулам (5), (6), считая, что G_+ — наличие генных мутаций у больных с ОДП и G_- — отсутствие генных мутаций. Будем распределять больных по группам с использованием вальдовского последовательного метода. Примем $\alpha = 0,01$ и $\beta = 0,01$. Тогда из (7), (8) следует, что прогностический порог для распределения больных в группу с мутациями равен 20; прогностический порог для распределения больных в группу без мутаций равен -20.

Пример 1. Больной № 22 из контрольной группы с генными мутациями.

Время до госпитализации — 132 часа, что соответствует 15 баллам.

Индекс тяжести состояния больного SOFA равен 8 баллам, что соответствует 12 баллам по прогностической таблице.

На втором шаге достигнут прогностический порог, равный 20 баллам, поэтому распределяем больного в группу с мутациями.

Пример 2. Больной № 38 из контрольной группы без генных мутаций.

Время до госпитализации — 24 часа, что соответствует -12 баллам.

Индекс SOFA равен 10 баллам, что соответствует 5 баллам по прогностической таблице.

Сроки формирования осложнений — 5 часов, что соответствует -12 баллам.

Индекс MODS равен 18 баллам, что соответствует -12 баллам по прогностической таблице.

На четвертом шаге достигнут прогностический порог -20, поэтому распределяем больного в группу без мутаций.

Все больные с генными мутациями из контрольной группы были правильно распределены в соответствующую группу. Для 10 больных без мутаций из контрольной группы для двух больных был получен неверный результат.

3.2. Степень тяжести ОСРП. Поскольку на этапе первоначальной сортировки пациентов главной задачей является учет всех тяжелобольных, будем принимать риск ошибок диагностики для пациентов с летальным исходом 1-й и 2-й групп тяжести, равным 0,05, а риск ошибок для пациентов с нелетальным исходом 3-й группы тяжести, равным 0,1.

Пациентов с ОСРП необходимо распределять в три группы, поэтому предлагаем использовать информационную меру дважды для отделения пациентов разных групп тяжести.

1. Тяжелобольных пациентов 1-й группы отделить от пациентов 2-й и 3-й групп. При этом в (5), (6) G_+ будет означать принадлежность пациента к 1-й группе тяжести, а G_- — ко 2-й или 3-й. Тогда согласно гипотезе H_0 пациент принадлежит к 1-й группе, согласно H_1 пациент принадлежит ко 2-й или 3-й группе. При этом ошибка первого рода α будет означать риск гиподиагностики, а ошибка второго рода β — риск гипердиагностики.

Прогностический порог для определения пациента в 1-ю группу находим по формуле

$$10 \lg \frac{1-\alpha}{\beta} = 10 \lg \frac{1-0,05}{0,05+0,1} \approx 8.$$

2. Пациентов 3-й группы, которые выжили, отделить от пациентов с летальным исходом 1-й и 2-й групп. В (5), (6) G_+ будет означать принадлежность пациента к 3-й группе тяжести, а G_- — к 1-й или 2-й. При этом смысл ошибок первого и второго рода изменится. Ошибка первого рода α будет означать риск гипердиагностики, а ошибка второго рода β — риск гиподиагностики.

Прогностическим порогом для определения больного в 3-ю группу является

$$10 \lg \frac{\alpha}{1-\beta} = 10 \lg \frac{0,1}{1-(0,5+0,5)} \approx 10.$$

Заметим, что поскольку местные лучевые поражения не были летальными только в случаях поражения меньше 30 % поверхности тела, информативность критериев в обоих случаях была разной: в первом случае, при определении больных 1-й группы, $J(R/G_+, R/G_-) = 2,75$, а $J(B/G_+, B/G_-) = 2,46$; во втором случае, при определении больных 3-й группы, $J(R/G_+, R/G_-) = 4,51$, а $J(B/G_+, B/G_-) = 1,07$.

Пример 3. Больной № 1004, $R = 11,8$; $B = 273,5$. Умер на 18-й день после облучения.

Вычисляем сумму баллов для проверки принадлежности пациента к 1-й группе тяжести: $R = 11,8$ соответствует 7 баллам; $B = 273,5$ соответствует 7 баллам.

Прогностический порог 8 достигнут, поэтому распределяем пациента в 1-ю группу тяжести.

Пример 4. Больной №1006, $R = 7,5$; $B = 42,4$. Умер на 84-й день после облучения.

Вычисляем сумму баллов для проверки принадлежности пациента к 1-й группе тяжести: $R = 7,5$ соответствует -3 баллам; $B = 42,4$ соответствует -1 баллу.

Прогностический порог 8 не достигнут, поэтому проверяем принадлежность пациента к 3-й группе: $R = 7,5$ соответствует -3 баллам; $B = 42,4$ соответствует 0 баллам.

Прогностический порог 10 не достигнут, поэтому и в 3-ю группу нельзя распределить пациента. Значит, пациент принадлежит ко 2-й группе тяжести.

Пример 5. Больной №1043, $R = 4$; $B = 10$. Облучение не привело к смерти больного.

Вычисляем сумму баллов для проверки принадлежности пациента к 1-й группе тяжести: $R = 4$ соответствует -5 баллам; $B = 10$ соответствует -5 баллам.

Прогностический порог 8 не достигнут, поэтому проверяем принадлежность пациента к 3-й группе: $R = 4$ соответствует 7 баллам; $B = 10$ соответствует 5 баллам.

Прогностический порог 10 достигнут, поэтому распределяем пациента в 3-ю группу тяжести.

Из 22 больных с летальным исходом 1-й и 2-й групп двое были распределены в 3-ю группу. Из 13 больных 3-й группы 9 были распределены во 2-ю группу больных без летального исхода.

4. КЛАСТЕРНЫЙ АНАЛИЗ

Под кластерным анализом будем понимать задачу разбиения исходных данных на группы таким образом, чтобы элементы, входящие в эти группы, были максимально похожи между собой по заранее заданному критерию. В настоящее время существует большое количество алгоритмов кластерного анализа, однако точная постановка задачи кластерного анализа до сих пор не сформулирована [17]. В значительной мере результат классификации зависит от выбранной меры близости объектов. Поэтому будем проводить анализ несколько раз, используя метрики Евклида и Воронина [17], и выбирать ту меру близости объектов, при которой степень адекватности модели окажется наибольшей.

Введем обозначения: n — число исследуемых больных, m — количество признаков-маркеров наличия осложнений, x_{ij} — j -й признак i -го пациента, $i = 1, n$; $j = 1, m$.

Будем проводить нормализацию показателей по формуле

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \quad (9)$$

где \bar{x}_j — среднее арифметическое j -го признака, σ_j — стандартное отклонение j -го признака.

4.1. Метод k -средних. Для каждого больного в евклидовом пространстве R^m будем рассматривать точку, координаты которой представляют значения m признаков-маркеров наличия осложнений. Пусть точка x_i соответствует i -му больному, а точка x_j соответствует j -му больному. Тогда расстояние между пациентами будем называть евклидовой метрикой [17]:

$$\rho(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik}^* - x_{jk}^*)^2}.$$

Кластерный анализ методом k -средних будем проводить по следующему алгоритму [18].

Шаг 1. Выбрать k наблюдений из числа n больных по одному из следующих правил:

- максимизируя начальное расстояние;
- случайным способом;
- выбирая первые k наблюдений.

Остальные $n - k$ объектов распределить по кластерам, считая выбранные k наблюдений центроидами кластеров.

Шаг 2. Вычислить центры кластеров как покоординатные средние. Перераспределить объекты по кластерам.

Шаг 3. Повторять шаг 2 до тех пор, пока не выполнится одно из следующих условий:

- кластерные центры стабилизировались, т.е. все объекты принадлежат тем же кластерам, что и на предыдущем шаге;
- число итераций равно заранее заданному максимальному значению.

4.1.1. Наличие генных модификаций у больных с ОДП. Проведем кластерный анализ методом k -средних, используя данные $n = 70$ больных с ОДП, у которых определено $m = 18$ признаков-маркеров наличия генных модификаций, показавших статистически достоверный результат дисперсионного анализа. Распределим больных по $k = 2$ кластерам.

Пример 6. Больной № 22 из контрольной группы с наличием генных мутаций. Расстояние от точки, которая соответствует данному больному, до центра первого кластера составляет $\rho(x_{22}, x^1) = 35,431$, до второго кластера — $\rho(x_{22}, x^2) = 5527,467$. Таким образом, больного № 22 относим к первому кластеру (с наличием генных мутаций).

Пример 7. Больной № 38 из контрольной группы без генных мутаций. Расстояние от точки, которая соответствует данному больному, до центра первого кластера составляет $\rho(x_{38}, x^1) = 8091,431$, до второго кластера — $\rho(x_{38}, x^2) = 93,182$. Таким образом, больного № 38 относим ко второму кластеру (без генных мутаций).

Из восьми больных контрольной группы с генными мутациями семь были правильно распределены в 1-й кластер, из 10 больных контрольной группы без генных мутаций восемь были правильно распределены во 2-й кластер.

4.1.2. Степень тяжести ОСРП. Будем проводить кластерный анализ методом k -средних с параметрами $n = 35$; $m = 2$; $k = 3$.

Пример 8. Больной № 1004, $R = 11,8$; $B = 273,5$. Умер на 18-й день после облучения. Вычислим расстояния, которые соответствуют данному больному, до центров кластеров: $\rho(x_{1004}, x^1) = 63,519$; $\rho(x_{1004}, x^2) = 236,916$; $\rho(x_{1004}, x^3) = 249,080$. Относим больного к первому кластеру.

Пример 9. Больной № 1006, $R = 7,5$; $B = 42,4$. Умер на 84-й день после облучения. Вычислим расстояния, которые соответствуют данному больному, до центров кластеров: $\rho(x_{1006}, x^1) = 167,632$; $\rho(x_{1006}, x^2) = 5,867$; $\rho(x_{1006}, x^3) = 18,052$. Относим больного ко второму кластеру.

Пример 10. Больной № 1043, $R = 4$; $B = 10$. Облучение не привело к смерти больного. Вычислим расстояния, которые соответствуют данному больному, до центров кластеров: $\rho(x_{1043}, x^1) = 200,127$; $\rho(x_{1043}, x^2) = 26,975$; $\rho(x_{1043}, x^3) = 14,551$. Относим больного к третьему кластеру.

Из 22 больных с летальным исходом четыре были распределены в «нелетальную» 3-ю группу, а один из 13 пациентов 3-й группы был распределен в «летальную» 2-ю группу.

4.2. Построение дендрограммы. Находим минимальное расстояние между кластерами в матрице расстояний: $\rho_{\min} = \min_{p=1, k, q=1, p-1} M(p, q) = M(p', q')$, где k

— количество кластеров на данном шаге. Объединяем кластеры p' , q' и пересчитываем матрицу расстояний M . Проведем кластерный анализ путем построения дендрограммы, вычисляя расстояние между пациентами с помощью метрики Воронина [17]:

$$\rho(x_i, x_j) = \frac{\sum_{k=1}^m c_k |x_{ik}^* - x_{jk}^*|}{m}, \quad (10)$$

где x_i, x_j — i -й и j -й пациенты соответственно, c_k — весовой коэффициент k -го признака.

Расстояние между кластерами S_l и S_k будем находить по методу средней связи [16]:

$$\rho(S_l, S_k) = \frac{1}{n_l n_k} \sum_{x_i \in S_l} \sum_{x_j \in S_k} \rho(x_i, x_j),$$

где n_l, n_k — объемы соответствующих кластеров.

Распределение больных по кластерам будем проводить по следующему алгоритму.

Шаг 1. Каждого больного распределяем в отдельный кластер. Строим M — матрицу расстояний между кластерами, которая является матрицей расстояний между пациентами, вычисленных по формуле (10).

Шаг 2 (и последующие шаги). Находим минимальное расстояние между кластерами в матрице расстояний: $\rho_{\min} = \min_{i=1, k, j=1, i-1} M(i, j) = M(i', j')$, где k —

количество кластеров на данном шаге.

Объединяем кластеры i' , j' и пересчитываем матрицу расстояний M между ними.

Шаг 3. Если количество кластеров k больше заранее заданного значения, повторяем шаг 2.

4.2.1. Наличие генных модификаций у больных с ОДП. Весовые коэффициенты c_k в (10) примем равными соответствующим информативностям (6), найденным ранее: $c_k = J(x_k / G_+, x_k / G_-)$, $k = 1, m$.

Полученное при этом распределение на кластеры аналогично распределению методом k -средних, что доказывает правильность наших рассуждений.

4.2.2. Степень тяжести ОСРП. Поскольку информативность признаков для распределения больных по трем группам однозначно определить нельзя, будем

находить весовые коэффициенты c_k для метрики Воронина из систем неравенств

$$\begin{cases} \rho(x_i, x_j) < 1, \text{ если } x_i, x_j \text{ имеют одинаковую степень тяжести,} \\ \rho(x_i, x_j) > 1, \text{ если } x_i, x_j \text{ имеют разную степень тяжести,} \\ c_1 > 0, c_2 > 0, \\ i, j = \overline{1, n}. \end{cases}$$

Суммируя значения с одинаковым знаком, получим систему неравенств

$$\begin{cases} 326,321c_1 + 253,185c_2 < 411, \\ 509,4c_1 + 497,936c_2 > 407, \\ c_1 > 0, c_2 > 0. \end{cases}$$

Одной из точек множества решений данной системы является точка $c = (0,5; 1)$.

В результате кластерного анализа с весовыми коэффициентами $c_1 = 0,5$ и $c_2 = 1$ был получен следующий результат: девять из 22 пациентов из 1-й и 2-й «летальных» групп были распределены в 3-ю «нелетальную» группу, все пациенты из 3-й группы были правильно распределены в один кластер.

5. КЛАССИФИКАЦИЯ МЕТОДОМ ПОВОРОТА СИСТЕМЫ КООРДИНАТ

Поскольку приведенные выше методы распределения пациентов по группам имеют высокий уровень ошибок для выборки малого размера (прогнозируемые степени тяжести у больных с ОСРП), предлагаем собственный альтернативный метод.

По формуле (9) нормализуем показатели и исключаем промахи из каждой группы. Для каждой группы найдем средние значения признаков-маркеров наличия осложнений. Получим точки — координаты средних для каждой группы:

$$x_i = (x_1^i, x_2^i, \dots, x_m^i), \quad i = \overline{1, k},$$

где m — количество признаков-маркеров, k — количество групп тяжести. Используем метод наименьших квадратов [13] для нахождения прямой, которая проходит через эти точки. В двухмерном случае уравнение прямой будет иметь вид $y = kx + b$. Повернем систему координат на угол $\alpha = \arctg k$ с помощью матрицы поворота

$$M = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}, \quad (11)$$

т.е. новые координаты будем вычислять по формулам

$$\begin{cases} x' = \cos \alpha \cdot x^* + \sin \alpha \cdot y^*, \\ y' = -\sin \alpha \cdot x^* + \cos \alpha \cdot y^*, \end{cases}$$

где $x^* = (x - \bar{x}) / \sigma_x$, $y^* = (y - \bar{y}) / \sigma_y$. После поворота системы координат группы пациентов распределяются так, что их можно разделить прямыми вида $x = \text{Const}$.

Для возвращения в первоначальную систему координат будем использовать формулы

$$\begin{cases} x^* = \cos \alpha \cdot x' - \sin \alpha \cdot y', \\ y^* = \sin \alpha \cdot x' + \cos \alpha \cdot y', \end{cases}$$

тогда

$$\begin{cases} x = x^* \sigma_x + \bar{x} = (\cos \alpha \cdot x' - \sin \alpha \cdot y') \sigma_x + \bar{x}, \\ y = y^* \sigma_y + \bar{y} = (\sin \alpha \cdot x' + \cos \alpha \cdot y') \sigma_y + \bar{y}. \end{cases}$$

5.1. Степень тяжести ОСРП. Применим описанный выше алгоритм. Нормализуем показатели и исключим промахи. В декартовой системе координат получим три точки — средние групп тяжести: (R_1^*, B_1^*) ; (R_2^*, B_2^*) ; (R_3^*, B_3^*) . По методу наименьших квадратов получим уравнение прямой, которая проходит через эти точки: $B^* = 0,8194R^* + 0,0002$. Повернем систему координат на угол $\alpha = \text{arctg } 0,8194$ с помощью матрицы поворота (11). Новые координаты будем вычислять по формулам

$$\begin{cases} R' = \cos \alpha \cdot R^* + \sin \alpha \cdot B^*, \\ B' = -\sin \alpha \cdot R^* + \cos \alpha \cdot B^*, \end{cases}$$

где $R^* = (R - \bar{R}) / \sigma_R$, $B^* = (B - \bar{B}) / \sigma_B$.

Распределение групп пациентов получим в новой системе координат. При этом координаты пациентов 3-й группы тяжести соответствуют условию $R' < -0,8$; 2-й группы — условию $-0,8 \leq R' \leq 0,8$; 1-й группы — условию $R' > 0,8$.

Возвратимся к первоначальной системе координат по формулам

$$\begin{cases} R^* = \cos \alpha \cdot R' - \sin \alpha \cdot B', \\ B^* = \sin \alpha \cdot R' + \cos \alpha \cdot B'. \end{cases}$$

Тогда

$$\begin{cases} R = R^* \sigma_R + \bar{R} = (\cos \alpha \cdot R' - \sin \alpha \cdot B') \sigma_R + \bar{R}, \\ B = B^* \sigma_B + \bar{B} = (\sin \alpha \cdot R' + \cos \alpha \cdot B') \sigma_B + \bar{B}. \end{cases}$$

Условие $R' < -0,8$ в начальных координатах будет иметь вид

$$\begin{aligned} \cos \alpha R^* + \sin \alpha B^* < -0,8; \\ R < -\text{tg } \alpha \frac{\sigma_R}{\sigma_B} B + \left(\frac{-0,8 \sigma_R}{\cos \alpha} + \bar{R} + \text{tg } \alpha \bar{B} \frac{\sigma_R}{\sigma_B} \right) \end{aligned}$$

или $R < -0,027B + 6,984$.

Условие $R' > 0,8$ в начальных координатах будет иметь вид

$$R > -\text{tg } \alpha \frac{\sigma_R}{\sigma_B} B + \left(\frac{0,8 \sigma_R}{\cos \alpha} + \bar{R} + \text{tg } \alpha \bar{B} \frac{\sigma_R}{\sigma_B} \right)$$

или $R > -0,027B + 13,106$.

Исходя из вышеизложенного, можно сформулировать алгоритм распределения пациентов по группам тяжести.

Будем распределять в 1-ю группу пациентов, для которых справедливо неравенство $R > -0,027B + 13,106$; больных, для которых справедливо условие $-0,027B + 6,984 \leq R \leq -0,027B + 13,106$, будем распределять во 2-ю группу; в 3-ю группу распределим пациентов, для которых справедливо неравенство $R < -0,027B + 6,984$.

Изобразим полученные результаты графически (рис. 1).

Пример 11. Больной № 1004, $R = 11,8$; $B = 273,5$. Умер на 18-й день после облучения. Для больного справедливо неравенство $R > -0,027B + 13,106$. Относим больного к первой группе.

Пример 12. Больной № 1006, $R = 7,5$; $B = 42,4$. Умер на 84-й день после облучения. Для больного справедливы неравенства $-0,027B + 6,984 \leq R \leq -0,027B + 13,106$. Относим больного ко второй группе.

Пример 13. Больной № 1043, $R = 4$; $B = 10$. Облучение не привело к смерти больного. Для больного справедливо неравенство $R < -0,027B + 6,984$. Относим больного к третьей группе.

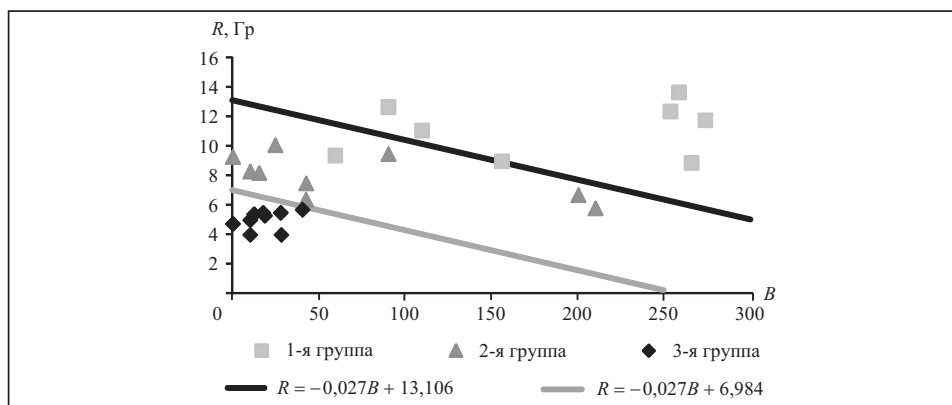


Рис. 1

Таблица 1

Способ прогнозирования	Показатели информативности прогностических методов (%) у пациентов с ОДП			
	Гипер-диагностика	Специфичность	Гипо-диагностика	Чувствительность
Построение прогностических таблиц	20	80	0	100
Кластерный анализ методом построения дендрограмм	20	80	12,5	87,5
Кластерный анализ методом <i>k</i> -средних	20	80	12,5	87,5

Таблица 2

Способ прогнозирования	Показатели информативности прогностических методов (%) у пациентов с ОСРП			
	Гипер-диагностика	Специфичность	Гипо-диагностика	Чувствительность
Построение прогностических таблиц	67	33	9	91
Кластерный анализ методом построения дендрограмм	0	100	41	59
Кластерный анализ методом <i>k</i> -средних	8	92	18	82
Поворот системы координат	23	77	14	86

После исключения промахов все пациенты были правильно распределены по соответственным группам тяжести. Проверим качество предположенной методики по всем данным. Трое из 22 пациентов с летальным исходом были распределены в 3-ю группу, трое из 13 пациентов без летального исхода — во 2-ю группу.

Результаты прогнозирования наличия генных модификаций у больных с ОДП различными методиками даны в табл. 1, а данные прогнозирования летальности и степени тяжести у пациентов с ОСРП различными методиками даны в табл. 2.

ЗАКЛЮЧЕНИЕ

Все применяемые в статье тесты относительно больных с ОДП одинаково специфичны. При этом построение диагностических таблиц с вальдовским последовательным анализом дало максимально чувствительный результат. Поэтому этот метод рекомендуется для использования на практике.

Самым чувствительным методом для больных с ОСПР оказался вальдовский последовательный анализ по построенным прогностическим таблицам. Однако этот метод имеет наименьшую специфичность. Поэтому для практического использования мы рекомендуем кластерный анализ методом k -средних и метод поворота системы координат.

Специфичность и чувствительность предложенного нами метода поворота системы координат находится на уровне классических методов кластерного анализа и вальдовского последовательного анализа, поэтому его можно предложить для использования на практике.

СПИСОК ЛИТЕРАТУРЫ

1. Савула М.М., Паучок В.К., Кравченко Н.С. Застосування модифікованого кластерного аналізу даних стандартного медичного обстеження для диференціальної діагностики деяких захворювань легень // Укр. пульмонологіч. журн. — 2010. — № 4. — С. 39–41.
2. Schaubel D.E. Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients // *Biostatistics*. — 2005. — 6, N 3. — P. 404–419.
3. Weatherall M., Shirtcliffe P., Travers J., Beasley R. Use of cluster analysis to define COPD phenotypes // *Eur. Respir. J.* — 2010. — 36. — P. 472–474.
4. Бородулин А.И., Свиридова А.В., Судаков О.В. Использование кластерного анализа в диагностике ишемической болезни сердца // *Приклад. информат. аспекты медицины*. — 2008. — 11, № 2. — С. 28–32.
5. Максим'юк В.В. Особливості клінічного перебігу гострого деструктивного панкреатиту в осіб із поліморфізмом R122H гена катіонного трипсिनогену (PRSS1) // *Буковин. медич. вісник*. — 2012. — 16, № 1 (61). — С. 45–48.
6. Акоев И.Г., Максимов Г.И., Малышев В.М. Лучевые поражения млекопитающих и статистическое моделирование. — М.: Атомиздат, 1976. — 96 с.
7. Математическая модель для прогноза летального исхода при кроветворной форме лучевого поражения / И.Б. Кеирим-Маркус, Н.Г. Даринская, В.Л. Гозенбук и др. // *Механизм системы взаимодействия при радиационном поражении. Теоретические предпосылки и модели*. — Пушкино: НЦБИ АН СССР, 1978. — С. 90–98.
8. Оценка неравномерности лучевой нагрузки при облучении человека в условиях ядерного взрыва / И.Б. Кеирим-Маркус, О.А. Кочетков, Е.Н. Чернов и др. // *Биологические эффекты неравномерных лучевых воздействий*. — М.: Атомиздат, 1974. — С. 5–9.
9. Близнюк А.П., Гозенбук В.П., Кеирим-Маркус И.Б. Дозиметрический критерий для кишечной формы острых лучевых поражений человека. Эмпирическая модель // *Радиобиология*. — 1988. — 28, вып. 6. — С. 809–811.
10. Altman D.G., Bland J.M. Diagnostic tests 1: sensitivity and specificity // *BMJ*. — 1994. — 308, June. — P. 1552.
11. Диагностика, клиническая картина и лечение острой лучевой болезни у пострадавших на Чернобыльской атомной электростанции. Сообщение 2. Некостномозговые синдромы лучевых поражений и их лечение / А.К. Гуськова, А.Е. Баранов, Г.Д. Селидовкин и др. // *Терапевтический архив* — 1989. — 12. — С. 99–103.
12. Бойчук Р.Р., Граціанська М.А. Математичний підхід до прогнозування виникнення кишкового синдрому та оцінки ступеня тяжкості гострого поєданого радіаційного ураження // *Буковин. медич. вісник*. — 2002. — 6, № 3. — С. 23–28.
13. Лакин Г.Ф. Биометрия: Учебное пособие для биологов спец. вузов. — 4-е изд., перераб. и доп. — М.: Высш. шк., 1990. — 352 с.
14. Кульбак С. Теория информации и статистика. — М.: Мир, 1967. — 408 с.
15. Гублер Е.В., Генкин А.А. Применение непараметрических критериев статистики в медико-биологических исследованиях. — Л.: Медицина, 1973. — 142 с.
16. Гублер Е.В. Вычислительные методы анализа и распознавания патологических процессов. — Л.: Медицина, 1978. — 350 с.
17. Мандель И.Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.
18. Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy cluster analysis. Methods for classification, data analysis and image recognition. — New York: John Wiley & Sons, 1999. — 289 p.

Поступила 21.11.2013