

Математическое моделирование и информатика

УДК 681.518.2:61+519.23/.24

В. А. Акулов

ДВУХКАСКАДНАЯ НЕПАРАМЕТРИЧЕСКАЯ МОДЕЛЬ И ЕЕ ПРИМЕНЕНИЕ В КАРДИОЛОГИИ

Предлагается математическая модель, получившая условное название «двухкаскадной», в качестве составной части новейших медицинских технологий, направленных на охрану индивидуального и коллективного здоровья. Основу модели составляют два каскада (группы формул), расположенных последовательно и выполняющих непараметрический анализ. В первом каскаде реализован двухвыборочный анализ с количественными данными, во втором - одновыборочный с номинальными данными, полученными в результате работы первого каскада. Модель адаптирована к проблемам суточного мониторинга артериального давления.

Современная практическая и теоретическая медицина выдвигает ряд задач, решение которых невозможно без двукратного выполнения непараметрического анализа. Это требует разработки и реализации в виде программного интерфейса весьма специфического класса математических моделей, которые условно названы двухкаскадными. В обобщенном виде подобное моделирование заключается в следующем.

Имеются некоторые объекты (j), состояние которых характеризуется количественными данными, представленными в виде вектор - столбцов:

$$X_{(j)}^{(k)} = \{p_{i1}^{(k)}, p_{i2}^{(k)}, \dots, p_{in}^{(k)}, \dots, p_{im}^{(k)}\}^T, \quad (1)$$

где X - измеряемый параметр; j - учетный номер пациента, $j = 1, 2, \dots, m$; k - номер исследования (мониторирования), $k = 1, 2, \dots, l$; p - значение измеряемого параметра; t - время; i - число зарегистрированных точек, $i = 1, 2, \dots, n$; T - символ транспонирования.

Примером (1) могут служить результаты серий регистраций медицинских показателей, разделенных по времени (схема «до- после»). Из множества (1) формируются группы по некоторым медицинским признакам.

Для каждой из групп необходимо решить две задачи по выявлению скрытых закономерностей, характеризующих как индивидуальные, так и групповые свойства объектов. Первая задача состоит в оценке однородности каждого из объектов по отношению к популяции, из которой они извлечены (см. п. 2). Вторая задача решается после первой и состоит в оценке значимости отличий внутри группы.

Как правило, решение перечисленных задач требует привлечения непараметрических методов анализа: малые и умеренные выборки для количественных данных, пропущенные значения и засорения; номинальные данные в задаче № 2.

В итоге формализация медицинских задач приводит к расчетной схеме, включающей в себя два последовательно расположенных каскада, каждый из которых выполняет непараметрический анализ, что и определило название модели. Важно отметить, что потребность в разработке подобного класса моделей обусловлена прогрессом медицинских технологий, связанных с охраной личного и коллективного здоровья.

Ниже рассмотрен вариант двухкаскадной модели (ДМ) применительно к задаче обработки результатов суточного мониторинга артериального давления (АД). Совершенно очевидно, что диапазон применения ДМ существенно шире и охватывает самые различные разделы практической, экспериментальной и теоретической медицины. Выбор контроля АД в качестве иллюстрации обусловлен, во-первых, наглядностью задачи и, во-вторых, это область первоочередного внедрения ДМ.

1. Суточное мониторирование АД и непараметрические модели. Одним из наиболее эффективных способов контроля АД является суточное мониторирование, которое заключается в измерении систо-

лического и диастолического давлений (Sys), (Dia), а также частоты сердечных сокращений (Pul) в заранее запрограммированные моменты времени или по командам пациента. В результате измерений формируется множество вектор-столбцов $\{X_{(j)}^{(k)}\}$ (1).

Современные системы оснащены разнообразными средствами визуализации и обработки данных. Среди них таблицы, графики, в том числе сглаженные кривые, тренды, гистограммы и т.п. («Detailed Analysis with a Personal Computer»). В то же время обращает на себя внимание отсутствие средств для научно обоснованного сопоставительного анализа. Это делает невозможной аргументированную оценку взаимных смещений распределений, описывающих состояние пациентов (в индивидуальном порядке и в группе), что серьезно затрудняет объективную оценку динамики (изменений по времени) жизненно важных показателей. В итоге, совершенно естественно формируется типичная для ДМ ситуация.

2. Постановка задачи. Предлагается постановка и решение двух задач, связанных с упоминавшимися проблемами индивидуального и коллективного здоровья.

Задача 1. Сравнительная оценка двух распределений на предмет выявления значимого смещения АД (анализ по схеме «до- после»).

Исходными данными являются результаты двух мониторингов одного и того же пациента, разделенных некоторым временным промежутком (серия измерений с номерами k и r):

$$X_{(j)}^{(k)} = \{p_{i1}^{(k)}, p_{i2}^{(k)}, \dots, p_{ii}^{(k)}, \dots, p_m^{(k)}\}^T; X_{(j)}^{(r)} = \{p_{i1}^{(r)}, p_{i2}^{(r)}, \dots, p_{ii}^{(r)}, \dots, p_m^{(r)}\}^T, \quad (2)$$

где $j = \text{Const}$; $k \neq r$; $X = \{\text{Sys}, \text{Dia}, \text{Pul}\}$.

Результатом решения задачи 1 является:

- выявление двух подгрупп векторов как моделей пациентов по признаку: «есть- нет» статистически значимый эффект смещения начального и конечного распределений АД по времени (предполагается, что оба мониторинга выполняются при близких схемах жизненной активности пациента);
- для подгруппы с выявленным эффектом смещения указывается его направление (положительная или отрицательная динамика)

$$X_{(j)}^{(r)} > (<) X_{(j)}^{(k)}.$$

Прикладные вопросы (проблемы индивидуального здоровья):

- индивидуализированный подбор стратегии и тактики лечения;
- количественная оценка эффективности проведенного лечения для каждого обследованного пациента;
- контроль стабильности АД в течение некоторого временного интервала: существование (отсутствие) положительной (отрицательной) динамики АД (Sys, Dia, Pul).

Задача 2. Оценка эффекта смещения АД в анализируемой группе пациентов, сформированной по некоторым медицинским признакам (пол, возраст, стадия заболеваний и т. п.).

Исходными данными являются результаты решения задачи 1, представленные в номинальной шкале.

Прикладные вопросы (проблемы коллективного здоровья):

- оценка качества проведенного лечения для выбранной группы пациентов,
- оценка стабильности АД для групп риска и здоровых пациентов.

3. Математические модели сопоставительного анализа (задачи 1 и 2). Рассмотрим проблему двувыворочного непараметрического анализа (каскад 1). Для решения задачи 1 воспользуемся моделью парного мониторинга (2). Специфика суточного мониторинга такова, что распределения не являются нормальными, а число измеряемых точек (i) невелико и составляет несколько десятков (см. п. 1). В итоге медицинская задача сравнения двух мониторингов сводится к задачам непараметрического анализа парных данных для случая количественной шкалы измерений. Из непараметрических критериев наиболее простым и наглядным является критерий знаков [1]. Применим его к задаче. Перейдем от парных выборок (2) к их разностям и представим их в матричном виде

$$Z_{(j)}^{(K)} = X_{(j)}^{(r)} - X_{(j)}^{(k)}. \quad (3)$$

Запишем (3) в форме аддитивной модели:

$$Z_{(j)}^{(K)} = \Theta_{(j)}^{(K)} + e_{(j)}^{(k)}, \quad (4)$$

где $\Theta_{(j)}^{(k)}$ - неслучайная величина (эффект смещения); $e_{(j)}^{(k)}$ - случайная нормально распределенная величина, имеющая нулевую медиану.

Формула (4) позволяет легко сформулировать нулевую (H_0) и альтернативную (H_1) гипотезы.

$H_0: \Theta = 0 \Rightarrow$ Эффект смещения отсутствует, выборки однородные, т.е. извлечены из популяций с равными медианами.

$H_1: \Theta \neq 0 \Rightarrow$ Выборки имеют статистически значимые отличия и извлечены из разных популяций

Если справедлива гипотеза об однородности выборок H_0 , то имеет место соотношение

$$P(X^{(r)} < X^{(k)}) = P(X^{(r)} > X^{(k)}) = 0,5,$$

где P^* - вероятность соответствующего события (*).

Следовательно, число положительных и отрицательных разностей Z_i , обозначим их соответственно через S_+ и S_- , не должно значительно отличаться от математического ожидания, равного $n/2$, т.е. $S_+ \approx S_-$. В итоге, моделью мониторингования становится широко известная схема Бернулли для случая $P = 0,5$, а статистический вывод (интерпретация парных данных (2)) принимает вид

$$\text{Max}\{S_+, S_-\} < S_{кр} \Rightarrow H_0; \quad (5)$$

$$\text{Max}\{S_+, S_-\} \geq S_{кр} \Rightarrow H_1. \quad (6)$$

Критическое значение $S_{кр}$ находится из таблиц (см., например [2]). Причем, $S_{кр} = S_{кр}(\alpha, (n - r))$, где α - уровень значимости (вероятность отвергнуть гипотезу, когда она верна), а r - число нулевых разностей Z_i .

Осталось подсчитать суммы S_+ , S_- и r :

$$S_{+(-)} = \sum_{i=1}^n \begin{cases} 1, \text{И} - \text{И} z_i > 0, (z_i < 0); \\ 0, \text{И} - \text{И} z_i < 0, (z_i > 0); \end{cases} \quad (7)$$

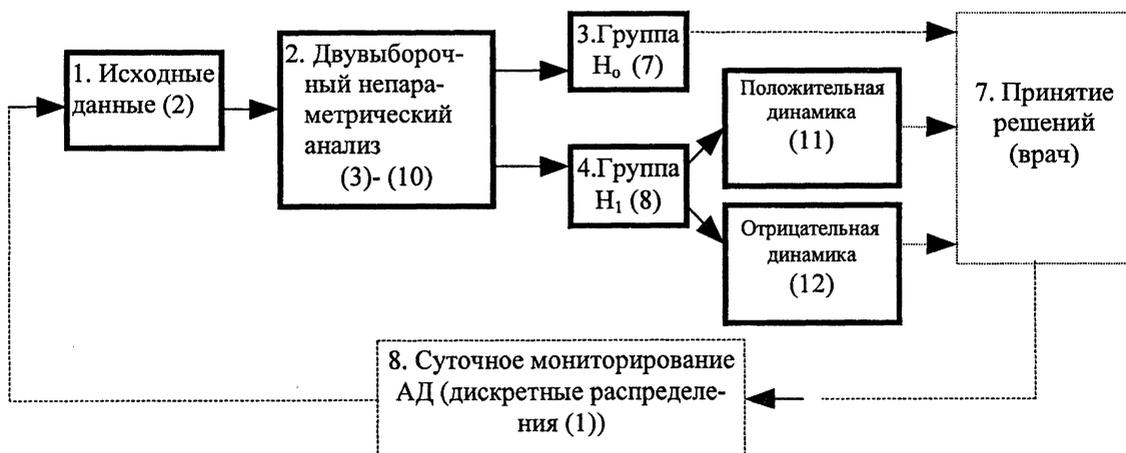
$$r = n - S_+ - S_- . \quad (8)$$

Если в результате анализа обнаружен эффект смещения, то определяется его направление:

$$S_+ > S_- \Rightarrow X^{(r)} > X^{(k)} \Rightarrow \text{«Положительная» динамика}, \quad (9)$$

$$S_+ < S_- \Rightarrow X^{(r)} < X^{(k)} \Rightarrow \text{«Отрицательная» динамика}. \quad (10)$$

Обобщим материалы, изложенные выше, и представим их в виде рисунка.



Функциональная схема 1-го каскада ДМ (в скобках указаны номера формул)

Суммируем результаты:

- формулы (2) - (10) образуют единый функциональный каскад, состоящий из 6 блоков;

- исходными данными являются дискретные распределения, измеренные в количественной шкале;
- выполняется двухвыборочный непараметрический анализ (блок 2);
- для случая H_1 (блок 4) определяется направление смещения распределений (блоки 5, 6);
- выявленные в ходе моделирования закономерности, связанные с индивидуальным здоровьем пациента, облегчают принятие решения (блок 7), в частности, уточняется программа дальнейших исследований (блок 8);
- модель отличается простотой, наглядностью, удобством реализации в программном интерфейсе.

Рассмотрим теперь проблему одновыборочного непараметрического анализа (каскад 2). Пусть имеется группа, состоящая из m пациентов. Подсчитаем число исходов с результатом «Да» (есть эффект смещения) и «Нет». Полученные результаты (атрибутивные данные) поместим в таблицу.

Атрибутивные данные для внутригруппового анализа

Число пациентов с результатом «да» (см. (8))	Число пациентов с результатом «нет» (см. (7))	Всего пациентов с парным мониторингом
m_1	m_2	$m = m_1 + m_2$

В общем случае $m_1 \neq m_2$. При этом возможны две существенно различные ситуации с точки зрения статистического вывода:

H_0 : имеющиеся отличия АД в анализируемой группе носят чисто случайный характер;

H_1 : имеющиеся отличия объясняются действием одного или нескольких факторов и имеет место эффект смещения.

В итоге медицинская задача сводится к одновыборочной задаче с номинальными исходными данными [2], что требует повторного применения непараметрического анализа (каскад № 2). Воспользуемся биномиальным критерием (в случае, если $n > 50$, следует применить критерий X^2). Статистический вывод примет вид

$$\text{Max}\{m_1, m_2\} < S_{\text{кр}2} \Rightarrow H_0; \quad (11)$$

$$\text{Max}\{m_1, m_2\} \geq S_{\text{кр}2} \Rightarrow H_1. \quad (12)$$

Величина $S_{\text{кр}2}$, как и S в формулах (5- 6), находится из таблиц, но $S_{\text{кр}} \neq S_{\text{кр}2}$.

4. *Программный интерфейс.* Совершенно очевидно, что двухкаскадные непараметрические алгоритмы требуют реализации в виде программного интерфейса. Наиболее важным требованием к нему является «непрерывность», под которой понимается выполнение расчетов в требуемой последовательности, начиная от анализа и корректировки исходных данных в соответствии с требованиями, предъявляемыми расчетными процедурами, и заканчивая получением корректных результатов.

В настоящее время разрабатывается опытный экземпляр такого интерфейса. В качестве базовой применена среда Excel как наиболее распространенная в медицинских приложениях. В то же время, следует отметить, что поставленные задачи могут быть решены с помощью специализированных статистических программ. При этом следует принимать во внимание следующее. Во-первых, двухкаскадность в них не предусмотрена, и каждое решение потребует формирования заданной последовательности вычислительных процедур. Во-вторых, не предусмотрен всеобъемлющий контроль исходных данных исходя из специфики задачи, что потребует разработки процедур подготовки исходных данных к расчету. В- третьих, отсутствует возможность изменения структуры информационных экранов в соответствии с требованиями весьма специфического пользователя, каковым является врач. В- четвертых, избыточность для данного приложения, а, следовательно, высокая стоимость. В этой связи отметим, что наиболее удобным и доступным является пакет Stadia [3].

Приведенные материалы позволяют сформулировать следующие выводы.

1. Разработана и апробирована весьма специфическая математическая модель, получившая название двухкаскадной (ДМ). Модель предназначена для применения в качестве составной части медицинских технологий, связанных с охраной индивидуального и коллективного здоровья. Основу модели составляет дважды выполняемый непараметрический анализ данных, организованный в виде двух последовательных каскадов (это и определило название).

2. Оба каскада связаны между собой и работают по последовательной схеме. Первый каскад оперирует с количественными данными, представленными в виде множества вектор- столбцов, и выполняет двухвыборочный непараметрический анализ на основе критерия знаков. Результатом моделирования является аргументированная оценка значимости имеющихся отличий анализируемых пар

вектор- столбцов. Как правило, это индивидуальные данные обследований пациентов на некотором временном интервале. По результатам анализа («есть- нет» эффект смещения), пациенты распределяются на две качественно различные группы, что влияет на выбор медицинской стратегии и тактики.

3. Полученные данные, измеренные в номинальной шкале, группируются по некоторым медицинским признакам и обрабатываются повторно (непараметрический одновыборочный анализ) на предмет оценки значимости отличий в новой группировке (каскад № 2).

4. Модель адаптирована к проблемам выявления скрытых закономерностей суточного мониторирования АД. В этом варианте первый каскад решает задачу оценки стабильности АД пациента на некотором временном интервале. В конечном счете решается задача существования или отсутствия положительной (отрицательной) динамики АД (Sys, Dia, Pul). Это важно для оценки состояний здоровых пациентов и групп риска, а также индивидуализированного подбора стратегии и тактики лечения больных. Второй каскад решает аналогичные проблемы, но применительно к группам пациентов, сформированным по некоторым медицинским критериям (пол, возраст, стадия заболеваний, методика лечения и т. п.).

5. Разработка специализированного интерфейса осуществляется в среде Excel. Выбор Excel обусловлен его распространенностью в медицинских приложениях. Показано, что поставленные задачи можно решить с помощью универсальных статистических пакетов. Перечислены затруднения, связанные с применением таких пакетов к специфическим задачам ДМ. Рекомендован пакет Stadia 6.0 автора А. П. Кулаичева.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере / Под ред. В. Э. Фигурнова. М.: ИНФРА-М, Финансы и статистика, 1995. 384 с.
2. *Рунион Р.* Справочник по непараметрической статистике: Современный подход. М.: Финансы и статистика, 1982. 198 с.
3. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows Stadia 6.0. М.: Информатика и компьютеры, 1996. 255 с.